CLOUD10

GENOMIC MEDICINE

LCR16a
LCR16v
LCR16c
LCR16l
LCR16u
LCR16b
LCR16k
13.3
13.2
13.1
12
11.2
11.1
11.1
11.2
12.1
12.2
13
21
22
23

Dr. T.J. Greer, MDPhD

This anthological textbook is composed of over 50 scholarly articles focused on topics and subtopics in genomic medicine. However, it begins with a brief synopsis of genetics, medicine, and genomic medicine. Please enjoy— Dr. T.J. Greer, MD/PhD, Harvard University

Introduction

DNA (deoxyribonucleic acid) is a complex molecule that contains all of the information necessary to build and maintain an organism. All living things have DNA within their cells. In fact, nearly every cell in a multicellular organism possesses the full set of DNA required for that organism. However, DNA does more than just specify the structure and function of living things — it also serves as the primary unit of heredity in organisms of all types. In other words, whenever organisms reproduce, a portion of their DNA is passed along to their offspring. This transmission of all or part of an organism's DNA helps ensure a certain level of continuity from one generation to the next, while still allowing for slight changes that contribute to the diversity of life (Nature Education, 2014). Moving on, medicine is a discipline centered on health and healing. It includes nurses, doctors, and various specialists. It covers diagnosis, treatment, and prevention of disease, medical research, and many other aspects of health. Medicine aims to promote and maintain health and wellbeing. Standard medicine is often called allopathic medicine. It involves the use of drugs or surgery, often supported by counseling and lifestyle measures. Alternative and complementary types of medicine include acupuncture, homeopathy, herbal medicine, art therapy, traditional Chinese medicine, and many more (Biggers, 2018). Combining these concepts, we see that genomic medicine is a discipline that involves using a patient’s genomic information as part of their clinical care. Other similar terms include individualized medicine, personalized medicine and precision medicine. For some conditions, genomic information can be used to help diagnose disease, predict outcomes and guide the appropriate treatment.

Now, let’s look at the central dogma of biology. DNA is the body’s blueprint. RNA is a cheaper copy of the DNA blueprint. Amino acids are tools, and proteins are the materials used to build the final products, products like cell walls, tissues, organelles, etc. (DNA replication): this process occurs during cell divisions and allows genetic information to pass from a mother to a daughter cell. DNA is copied to RNA (RNA transcription): this is the first step of the gene expression process. mRNA molecules are responsible for the transfer of the genetic information from the nucleus (where the information is stored) to the cytoplasm (where protein synthesis occurs). RNA is then used as a template to synthetize proteins (protein translation): this is the final step of the gene expression process. The ribosome, a specialized complex that "reads" the sequence of mRNA nucleotides, interacts with the mRNA. Each sequence of three nucleotides, called a codon, typically codes for one specific amino acid. The protein is assembled one amino acid at a time by tRNA. Protein assembly continues until the ribosome encounters a “stop” codon–a sequence of three nucleotides that does not code for an amino acid (Martinez, 2022). With this foundation laid, let’s further contextualize medical genomics. Medicine’s history is full of moments when one discovery suddenly redefined what was possible. From antibiotics to artificial intelligence, these breakthroughs haven't just changed how we treat disease—they've completely reshaped how we understand the human body.

Next, let's look at ten of the most transformative innovations in medicine, followed by a few modern breakthroughs that continue to contextualize genomic medicine. First, few discoveries have saved as many lives as antibiotics. When Alexander Fleming discovered penicillin in 1928, he couldn't have known it would change the course of medical history (Fleming, 1929). Before antibiotics, infections like pneumonia or sepsis were often fatal. Mass production during World War II made penicillin widely available, saving thousands of soldiers from infected wounds (Lax, 2004). Later antibiotics like streptomycin and tetracycline expanded the fight against tuberculosis and other diseases (Waksman, 1953). Although antibiotic resistance is now a growing threat (Ventola, 2015), the discovery remains one of humanity's greatest scientific victories.

Next, vaccinations transformed medicine from reactive to preventive care. Edward Jenner's smallpox vaccine in 1796 proved that exposure to a harmless form of a virus could protect against deadly infection (Jenner, 1798). The result? Smallpox became the first human disease ever eradicated in 1980 (Fenner et al., 1988). Since then, vaccines have nearly eliminated polio and drastically reduced measles, diphtheria, and tetanus (Plotkin, 2014). The success of the COVID-19 mRNA vaccines also showed how rapidly biotechnology can respond to new threats (Krammer, 2020). Vaccines continue to be the backbone of global public health.

Moving on, before anesthesia, surgery was brutal and often deadly. That changed in 1846 when dentist William Morton publicly demonstrated the use of ether at Massachusetts General Hospital (Snow, 1847). For the first time, patients could undergo surgery without agony. The development of chloroform and later safer inhalation agents made complex procedures possible and humane (Bulloch, 1947). Modern anesthesia combines powerful drugs with precise monitoring, making everything from heart transplants to cesarean deliveries possible (Miller, 2010). It's hard to imagine modern medicine without this ability to control pain and consciousness.

Next, when Wilhelm Conrad Röntgen discovered X-rays in 1895, he unlocked a new way to look inside the human body without surgery (Röntgen, 1896). This discovery launched the entire field of medical imaging. Today, CT scans, MRIs, and ultrasounds allow doctors to detect tumors, broken bones, and internal injuries with incredible accuracy (Bushberg & Boone, 2011). Imaging isn't just diagnostic—it's predictive, allowing early detection of disease before symptoms appear. From cancer screening to prenatal care, imaging has become one of the most indispensable tools in medicine.

Furthermore, before the 19th century, doctors believed diseases were caused by "miasmas" or bad air. Louis Pasteur and Robert Koch changed everything when they proved that microorganisms were responsible for infection (Pasteur, 1861; Koch, 1876). Germ theory led to cleaner hospitals, safer surgeries, and ultimately, antiseptics and antibiotics. This shift in understanding did more than save lives—it created the foundation for all modern medicine. As historian Bruno Latour noted, it changed how we think about responsibility and hygiene in everyday life (Latour, 1988). Germ theory is the cornerstone upon which nearly all other breakthroughs rest.

Next, replacing a failing organ with a healthy one was once considered science fiction. That changed in 1954 when Joseph Murray performed the first successful kidney transplant between identical twins (Murray et al., 1955). Advances in immunosuppressive drugs like cyclosporine made organ transplants more widely possible (Calne et al., 1978). Liver, heart, and even lung transplants are now routine in many countries. Organ transplantation doesn't just save lives—it also represents an incredible collaboration between surgery, immunology, and ethics (Starzl, 1963). The continuing shortage of donor organs keeps driving research into artificial and lab-grown alternatives.

Moreover, the discovery of DNA's double helix by Watson and Crick in 1953 paved the way for genetic engineering (Watson & Crick, 1953). The ability to manipulate genes reached a new level with the invention of CRISPR-Cas9 in 2012, allowing precise editing of DNA sequences (Doudna & Charpentier, 2014). Genetic engineering has enormous potential for curing inherited diseases, developing new drugs, and improving crops. But it also raises ethical questions about human enhancement and genetic inequality. Still, it's one of the most promising frontiers in modern medicine.

Moving on, for centuries, the heart was seen as too sacred—and too dangerous—to operate on. That changed in the 20th century with advances like the heart-lung machine and coronary bypass surgery. In 1967, Christiaan Barnard performed the first human heart transplant, proving that cardiac surgery could replace a failing heart entirely (Barnard, 1967). Techniques like angioplasty, valve repair, and minimally invasive procedures now save millions of lives every year (Braunwald, 2015). Heart surgery represents the best of modern medicine: skill, courage, and constant innovation.

Next, Joseph Lister revolutionized surgery in the 1860s by introducing carbolic acid to sterilize instruments and wounds (Lister, 1867). Mortality rates from surgery plummeted almost overnight. Antiseptic practices, combined with sterile operating environments, made hospitals far safer and childbirth much less risky (Hauser, 2013). This idea—clean hands, clean tools, clean spaces—seems obvious today, but it marked a turning point in medicine. Without antiseptics, even minor surgeries could still be fatal.

Finally, before the discovery of insulin, diabetes was like a death sentence. In 1921, Frederick Banting and Charles Best isolated insulin from the pancreas and proved it could control blood sugar in diabetic patients (Banting & Best, 1922). Their discovery transformed diabetes from a terminal disease into a manageable condition. Today, biosynthetic insulin and advanced delivery systems like insulin pumps continue to improve patients' quality of life (Bliss, 1982). Insulin is a perfect example of how basic research can turn into a lifesaving treatment used worldwide.

While those ten breakthroughs changed the past, new technologies are shaping the future. Epidemiology, the study of disease patterns, helps control outbreaks and pandemics. John Snow's work tracing cholera in 1854 laid the foundation for modern public health (Snow, 1855). 3D printing now allows doctors to create custom prosthetics and even bio-printed organs (Murphy & Atala, 2014). Meanwhile, artificial intelligence (AI) is transforming diagnostics, from detecting cancer in scans to predicting heart disease risk (Topol, 2019). These innovations suggest that the next revolution in medicine might come not from a single discovery, but from how technologies work together.

Finally, there are many potential applications of genomic medicine. 1. Examination of variation among healthy individuals 2. Understanding disease risk, susceptibility, and etiology 3. Disease prevention 4. Early diagnosis at the preclinical stage (to help modify disease course, e.g., diet and exercise in diabetes) 5. Identification of new disease-related mutations 6. Accurate diagnosis of challenging cases with inconclusive results for clinical parameters 7. Accurate disease classification based on molecular signature 8. Health management (prognosis and predictive markers) 9. Identification of personal drug-related profile 10. Selection of patients for clinical trials 11. Monitoring disease status (e.g., recurrence) 12. Monitoring tumor evolution in response to treatment 13. Developing new targeted therapies, especially for metastatic tumors refractory to treatment

Strategies to increase the cost-effectiveness of genomic screening strategies. 1. Target screening to a high-risk population 2. Focus on chronic diseases 3. Focus screening on diseases with known interventions 4. Identify molecular changes directly related to short-term benefits 5. Limit screening to diseases with a higher prevalence 6. Interpret data in a clinical context (e.g., environmental factors)

National Human Genome Research Institute

Talking Glossary of Genomic and Genetic Terms

The glossary features nearly 250 terms explained in an easy-to-understand way by leading scientists and professionals at the National Human Genome Research Institute.

Each term includes a spoken definition and narrative, and most include an illustration to provide a visual explanation. For the full audio and visual experience, please visit genome.gov/tg

-A-

Adenine
Adenine (A) is one of the four nucleotide bases in DNA, with the other three being cytosine (C), guanine (G) and thymine (T). Within a double-stranded DNA molecule, adenine bases on one strand pair with thymine bases on the opposite strand. The sequence of the four nucleotide bases encodes DNA's information.

Allele
An allele is one of two or more versions of DNA sequence (a single base or a segment of bases) at a given genomic location. An individual inherits two alleles, one from each parent, for any given genomic location where such variation exists. If the two alleles are the same, the individual is homozygous for that allele. If the alleles are different, the individual is heterozygous.

Alternative Splicing
Alternative splicing is a cellular process in which exons from the same gene are joined in different combinations, leading to different, but related, mRNA transcripts. These mRNAs can be translated to produce different proteins with distinct structures and functions — all from a single gene.

Amino Acid
An amino acid is the fundamental molecule that serves as the building block for proteins. There are 20 different amino acids. A protein consists of one or more chains of amino acids (called polypeptides) whose sequence is encoded in a gene. Some amino acids can be synthesized in the body, but others (essential amino acids) cannot and must be obtained from a person's diet.

Aneuploidy
Aneuploidy is an abnormality in the number of chromosomes in a cell due to loss or duplication. In humans, aneuploidy would be any number of chromosomes other than the usual 46.

Animal Model
An animal model is a non-human species used in biomedical research because it can mimic aspects of a biological process or disease found in humans. Animal models (e.g., mice, rats, zebrafish and others) are sufficiently like humans in their anatomy, physiology or response to a pathogen that researchers can extrapolate the results of animal model studies to better understand human physiology and disease. By using animal models, researchers can perform experiments that would be impractical or ethically prohibited with humans.

Anticodon
A codon is a DNA or RNA sequence of three nucleotides (a trinucleotide) that forms a unit of genetic information encoding a particular amino acid. An anticodon is a trinucleotide sequence located at one end of a transfer RNA (tRNA) molecule, which is complementary to a corresponding codon in a messenger RNA (mRNA) sequence. Each time an amino acid is added to a growing polypeptide during protein synthesis, a tRNA anticodon pairs with its complementary codon on the mRNA molecule, ensuring that the appropriate amino acid is inserted into the polypeptide.

Antisense
Antisense is the non-coding DNA strand of a gene. In a cell, antisense DNA serves as the template for producing messenger RNA (mRNA), which directs the synthesis of a protein.

Autism
Autism is a condition related to brain development that can cause significant social, communication and behavioral challenges. Symptoms usually appear before the age of three. The exact cause of autism is not entirely known, although genetics clearly plays an important role. Autism is one of a group of related developmental conditions sometimes called the autism spectrum that affect people differently and to varying degrees.

Autosomal Dominant Disorder
Autosomal dominant is a pattern of inheritance characteristic of some genetic disorders. “Autosomal” means that the gene in question is located on one of the numbered, or non-sex, chromosomes. “Dominant” means that a single copy of the mutated gene (from one parent) is enough to cause the disorder. A child of a person affected by an autosomal dominant condition has a 50% chance of being affected by that condition via inheritance of a dominant allele. By contrast, an autosomal recessive disorder requires two copies of the mutated gene (one from each parent) to cause the disorder. Huntington’s disease is an example of an autosomal dominant genetic disorder.

Autosomal Recessive Disorder
Autosomal recessive is a pattern of inheritance characteristic of some genetic disorders. “Autosomal” means that the gene in question is located on one of the numbered, or non-sex, chromosomes. “Recessive” means that two copies of the mutated gene (one from each parent) are required to cause the disorder. In a family where both parents are carriers and do not have the disease, roughly a quarter of their children will inherit two disease-causing alleles and have the disease. By contrast, an autosomal dominant disorder requires only a single copy of the mutated gene from one parent to cause the disorder. Sickle cell anemia is an example of an autosomal recessive genetic disorder.

Autosome
An autosome is one of the numbered chromosomes, as opposed to the sex chromosomes. Humans have 22 pairs of autosomes and one pair of sex chromosomes (XX or XY). Autosomes are numbered roughly in relation to their sizes. The largest autosome — chromosome 1 — has approximately 2,800 genes; the smallest autosome — chromosome 22 — has approximately 750 genes.

-B-

Base Pair
A base pair consists of two complementary DNA nucleotide bases that pair together to form a “rung of the DNA ladder.” DNA is made of two linked strands that wind around each other to resemble a twisted ladder — a shape known as a double helix. Each strand has a backbone made of alternating sugar (deoxyribose) and phosphate groups. Attached to each sugar is one of four bases: adenine (A), cytosine (C), guanine (G) or thymine (T). The two strands are held together by hydrogen bonds between pairs of bases: adenine pairs with thymine, and cytosine pairs with guanine.

Bioinformatics
Bioinformatics, as related to genetics and genomics, is a scientific subdiscipline that involves using computer technology to collect, store, analyze and disseminate biological data and information, such as DNA and amino acid sequences or annotations about those sequences. Scientists and clinicians use databases that organize and index such biological information to increase our understanding of health and disease and, in certain cases, as part of medical care.

Birth Defect
A birth defect (also called a congenital defect) is a physical or physiological abnormality present in a baby at birth. Birth defects can be caused by genetic factors, prenatal events during pregnancy or a combination of both. Some birth defects are easy to see (such as an extra or missing finger), while others (such as an enzyme deficiency) are identified through special tests.

BRCA1/BRCA2

BRCA1 and *BRCA2* are the first two genes found to be associated with inherited forms of breast cancer and ovarian cancer. People with mutations in either *BRCA1* or *BRCA2* have a much higher risk for developing breast, ovarian or other types of cancer than those without mutations in the genes. Both BRCA1 and BRCA2 normally act as tumor suppressors, meaning they help to regulate cell division. Most people have two active copies of these genes. When one of the two copies becomes inactive due to an inherited mutation, a person's cells are left with only one copy. If this remaining copy also becomes inactivated, then uncontrolled cell growth results, which leads to breast, ovarian or other types of cancer.

-C-

Cancer

Cancer is a disease in which some of the body's cells grow uncontrollably. There are many different types of cancer, and each begins when a single cell acquires a genomic change (or mutation) that allows the cell to divide and multiply unchecked. Additional mutations can cause the cancer to spread to other sites. Such mutations can be caused by errors during DNA replication or result from DNA damage due to environmental exposures (such as tobacco smoke or the sun's ultraviolet rays). In certain cases, mutations in cancer genes are inherited, which increases a person's risk of developing cancer.

Cancer-Susceptibility Gene

A cancer-susceptibility gene is a gene that, when changed (or mutated), gives an individual an increased risk for developing cancer. Individuals who have inherited mutations in certain cancer-susceptibility genes have a lifetime risk of cancer that is significantly higher than the general population (e.g., *BRCA1/BRCA2*). Individuals in the high-risk category may benefit from more frequent cancer screens. There are also many gene variants associated with a small increase in risk. In some cases, environmental factors may also play a role.

Candidate Gene

The term candidate gene refers to a gene that is believed to be related to a particular trait, such as a disease or a physical attribute. Because of its genomic location or its known function, the gene is suspected to play a role in that trait, thus making it a candidate for additional study.

Carcinogen

A carcinogen is a substance, organism or agent capable of causing cancer. Carcinogens may occur naturally in the environment (such as ultraviolet rays in sunlight and certain viruses) or may be generated by humans (such as automobile exhaust fumes and cigarette smoke). Most carcinogens work by interacting with a cell's DNA to produce mutations.

Carrier

A carrier, as related to genetics, is an individual who "carries" and can pass on to its offspring a genomic variant (allele) associated with a disease (or trait) that is inherited in an autosomal recessive or sex-linked manner, and who does not show symptoms of that disease (or features of that trait). The carrier has inherited the variant allele from one parent and a normal allele from the other parent. Any offspring of carriers is at risk of inheriting a variant allele from their parents, which would result in that child having the disease (or trait).

Carrier Screening

Carrier screening involves testing to see if a person "carries" a genetic variation (allele) associated with a specific disease or trait. A carrier has inherited a normal and a variant allele for a disease- or trait-associated gene, one from each parent. Most typically, carrier screening is performed to look for recessively inherited diseases when the suspected carrier has no symptoms of the disease, but that person's offspring could have the disease if the other parent is a carrier of a harmful variant in the same gene.

Cell-Free DNA Testing

Cell-free DNA testing is a laboratory method that involves analyzing free (i.e., non-cellular) DNA contained within a biological sample, most often to look for genomic variants associated with a hereditary or genetic disorder. For example, prenatal cell-free DNA testing is a non-invasive method used during pregnancy that examines the fetal DNA that is naturally present in the maternal bloodstream. Cell-free DNA testing is also used for the detection and characterization of some cancers and to monitor cancer therapy.

Centimorgan (cM)

A centimorgan (abbreviated cM) is a unit of measure for the frequency of genetic recombination. One centimorgan is equal to a 1% chance that two markers on a chromosome will become separated from one another due to a recombination event during meiosis (which occurs during the formation of egg and sperm cells). On average, one centimorgan corresponds to roughly 1 million base pairs in the human genome.

Central Dogma

The central dogma of molecular biology is a theory stating that genetic information flows only in one direction, from DNA to RNA to protein (or RNA to protein).

Centromere

The centromere appears as a constricted region of a chromosome and plays a key role in helping the cell divide up its DNA during division (mitosis and meiosis). Specifically, it is the region where the cell's spindle fibers attach. Following attachment of the spindle fibers to the centromere, the two identical sister chromatids that make up the replicated chromosome are pulled to opposite sides of the dividing cell, such that the two resulting daughter cells end up with identical DNA.

Chromatid

A chromatid is one of the two identical halves of a chromosome that has been replicated in preparation for cell division. The two “sister” chromatids are joined at a constricted region of the chromosome called the centromere. During cell division, spindle fibers attach to the centromere and pull each of the sister chromatids to opposite sides of the cell. Soon after, the cell divides in two, resulting in daughter cells with identical DNA.

Chromatin

Chromatin refers to a mixture of DNA and proteins that form the chromosomes found in the cells of humans and other higher organisms. Many of the proteins — namely, histones — package the massive amount of DNA in a genome into a highly compact form that can fit in the cell nucleus.

Chromosome

Chromosomes are threadlike structures made of protein and a single molecule of DNA that serve to carry the genomic information from cell to cell. In plants and animals (including humans), chromosomes reside in the nucleus of cells. Humans have 22 pairs of numbered chromosomes (autosomes) and one pair of sex chromosomes (XX or XY), for a total of 46. Each pair contains two chromosomes, one coming from each parent, which means that children inherit half of their chromosomes from their mother and half from their father. Chromosomes can be seen through a microscope when the nucleus dissolves during cell division.

Cloning

Cloning, as it relates to genetics and genomics, involves using scientific methods to make identical, or virtually identical, copies of an organism, cell or DNA sequence. The phrase “molecular cloning” typically refers to isolating and copying a particular DNA segment of interest for further study.

Codominance

Codominance, as it relates to genetics, refers to a type of inheritance in which two versions (alleles) of the same gene are expressed separately to yield different traits in an individual. That is, instead of one trait being dominant over the other, both traits appear, such as in a plant or animal that has more than one pigment color.

Codon

A codon is a DNA or RNA sequence of three nucleotides (a trinucleotide) that forms a unit of genomic information encoding a particular amino acid or signaling the termination of protein synthesis (stop signals). There are 64 different codons: 61 specify amino acids and 3 are used as stop signals.

Complex Disease

A complex disease (or condition), when discussed in the context of genetics, reflects a disorder that results from the contributions of multiple genomic variants and genes in conjunction with significant influences of the physical and social environment. For this reason, complex diseases are also called multifactorial diseases. This stands in contrast to a “simple” genetic disease that is more directly caused by mutations in a single gene. Common examples of complex genetic diseases include heart disease, diabetes and cancer.

Congenital

Congenital refers to a condition or trait that exists at birth. Congenital conditions or traits may be hereditary or result from an action or exposure occurring during pregnancy or at birth, or they may be due to a combination of these factors.

Contig

A contig (as related to genomic studies; derived from the word “contiguous”) is a set of DNA segments or sequences that overlap in a way that provides a contiguous representation of a genomic region. For example, a clone contig provides a physical map of a set of cloned segments of DNA across a genomic region, while a sequence contig provides the actual DNA sequence of a genomic region.

Copy DNA (cDNA)

cDNA (short for copy DNA; also called complementary DNA) is synthetic DNA that has been transcribed from a specific mRNA through a reaction using the enzyme reverse transcriptase. While DNA is composed of both coding and non-coding sequences, cDNA contains only coding sequences. Scientists often synthesize and use cDNA as a tool in gene cloning and other research experiments.

Copy Number Variation (CNV)

Copy number variation (abbreviated CNV) refers to a circumstance in which the number of copies of a specific segment of DNA varies among different individuals’ genomes. The individual variants may be short or include thousands of bases. These structural differences may have come about through duplications, deletions or other changes and can affect long stretches of DNA. Such regions may or may not contain a gene(s).

CRISPR

CRISPR (short for “clustered regularly interspaced short palindromic repeats”) is a technology that research scientists use to selectively modify the DNA of living organisms. CRISPR was adapted for use in the laboratory from naturally occurring genome editing systems found in bacteria.

Crossing Over

Crossing over, as related to genetics and genomics, refers to the exchange of DNA between paired homologous chromosomes (one from each parent) that occurs during the development of egg and sperm cells (meiosis). This process results in new combinations of alleles in the gametes (egg or sperm) formed, which ensures genomic variation in any offspring produced.

Cystic Fibrosis (CF)

Cystic fibrosis (abbreviated CF) is a genetic disorder that causes mucus to build up in certain organs of the body, particularly the lungs and pancreas, resulting in breathing problems, respiratory infections and faulty digestion. Caused by a mutation in a single gene (called *CFTR*), the disorder is inherited as an autosomal recessive trait, meaning that an affected individual inherits two mutated copies of the gene, one from each parent. In the past, CF was almost always fatal in childhood. Today, however, with improvements in screening and treatments, individuals with CF may live into their 30s or 40s, or even longer.

Cytogenetics

Cytogenetics is a branch of biology focused on the study of chromosomes and their inheritance, especially as applied to medical genetics. Chromosomes are microscopic structures containing DNA that reside within the nucleus of a cell. During cell division, these structures become condensed and are visible with a microscope. Special staining techniques can be used to assess the number and structure of a person’s chromosomes as part of diagnostic testing. The number and/or structure of chromosomes is known to be altered in certain genetic diseases.

Cytosine

Cytosine (C) is one of the four nucleotide bases in DNA, with the other three being adenine (A), guanine (G) and thymine (T). Within a double-stranded DNA molecule, cytosine bases on one strand pair with guanine bases on the opposite strand. The sequence of the four nucleotide bases encodes DNA’s information.

-D-

Data Science
Data science involves the study of large, complex data sets that arise from various types of research projects. With respect to genomic studies, such work requires expertise in quantitative scientific disciplines such as bioinformatics, computational biology and biostatistics.

Deletion
A deletion, as related to genomics, is a type of mutation that involves the loss of one or more nucleotides from a segment of DNA. A deletion can involve the loss of any number of nucleotides, from a single nucleotide to an entire piece of a chromosome.

Deoxyribonucleic Acid (DNA)
Deoxyribonucleic acid (abbreviated DNA) is the molecule that carries genetic information for the development and functioning of an organism. DNA is made of two linked strands that wind around each other to resemble a twisted ladder — a shape known as a double helix. Each strand has a backbone made of alternating sugar (deoxyribose) and phosphate groups. Attached to each sugar is one of four bases: adenine (A), cytosine (C), guanine (G) or thymine (T). The two strands are connected by chemical bonds between the bases: adenine bonds with thymine, and cytosine bonds with guanine. The sequence of the bases along DNA's backbone encodes biological information, such as the instructions for making a protein or RNA molecule.

Diploid
Diploid is a term that refers to the presence of two complete sets of chromosomes in an organism's cells, with each parent contributing a chromosome to each pair. Humans are diploid, and most of the body's cells contain 23 chromosomes pairs. Human gametes (egg and sperm cells), however, contain a single set of chromosomes and are said to be haploid.

DNA Fingerprinting
DNA fingerprinting is a laboratory technique used to determine the probable identity of a person based on the nucleotide sequences of certain regions of human DNA that are unique to individuals. DNA fingerprinting is used in a variety of situations, such as criminal investigations, other forensic purposes and paternity testing. In these situations, one aims to "match" two DNA fingerprints with one another, such as a DNA sample from a known person and one from an unknown person.

DNA Replication
DNA replication is the process by which the genome's DNA is copied in cells. Before a cell divides, it must first copy (or replicate) its entire genome so that each resulting daughter cell ends up with its own complete genome.

DNA Sequencing
DNA sequencing refers to the general laboratory technique for determining the exact sequence of nucleotides, or bases, in a DNA molecule. The sequence of the bases (often referred to by the first letters of their chemical names: A, T, C, and G) encodes the biological information that cells use to develop and operate. Establishing the sequence of DNA is key to understanding the function of genes and other parts of the genome. There are now several different methods available for DNA sequencing, each with its own characteristics, and the development of additional methods represents an active area of genomics research.

Dominant Traits and Alleles
Dominant, as related to genetics, refers to the relationship between an observed trait and the two inherited versions of a gene related to that trait. Individuals inherit two versions of each gene, known as alleles, from each parent. In the case of a dominant trait, only one copy of the dominant allele is required to express the trait. The effect of the other allele (the recessive allele) is masked by the dominant allele. Typically, an individual who carries two copies of a dominant allele exhibits the same trait as those who carry only one copy. This contrasts to a recessive trait, which requires that both alleles be present to express the trait.

Double Helix

Double helix, as related to genomics, is a term used to describe the physical structure of DNA. A DNA molecule is made up of two linked strands that wind around each other to resemble a twisted ladder in a helix-like shape. Each strand has a backbone made of alternating sugar (deoxyribose) and phosphate groups. Attached to each sugar is one of four bases: adenine (A), cytosine (C), guanine (G) or thymine (T). The two strands are connected by chemical bonds between the bases: adenine bonds with thymine, and cytosine bonds with guanine.

Down Syndrome (Trisomy 21)

Down syndrome (also called Trisomy 21) is a genetic condition caused by an error in the process that replicates and then divides up the pairs of chromosomes during cell division, resulting in the inheritance of an extra full or partial copy of chromosome 21 from a parent. This extra chromosomal DNA causes the intellectual disabilities and physical features characteristic of Down syndrome, which vary among individuals.

Duplication

Duplication, as related to genomics, refers to a type of mutation in which one or more copies of a DNA segment (which can be as small as a few bases or as large as a major chromosomal region) is produced. Duplications occur in all organisms. For example, they are especially prominent in plants, although they can also cause genetic diseases in humans. Duplications have been an important mechanism in the evolution of the genomes of humans and other organisms.

-E-

Electrophoresis

Electrophoresis is a laboratory technique used to separate DNA, RNA or protein molecules based on their size and electrical charge. An electric current is used to move the molecules through a gel or other matrix. Pores in the gel or matrix work like a sieve, allowing smaller molecules to move faster than larger molecules. To determine the size of the molecules in a sample, standards of known sizes are separated on the same gel and then compared to the sample.

Environmental Factors

Environmental factors, as related to genetics, refers to exposures to substances (such as pesticides or industrial waste) where we live or work, behaviors (such as smoking or poor diet) or stressful situations (such as racism) that can increase an individual’s risk of disease. Genetic studies often take environmental factors into consideration, as these exposures can increase an individual’s risk of genetic damage or disease.

Epigenetics

Epigenetics (also sometimes called epigenomics) is a field of study focused on changes in DNA that do not involve alterations to the underlying sequence. The DNA letters and the proteins that interact with DNA can have chemical modifications that change the degrees to which genes are turned on and off. Certain epigenetic modifications may be passed on from parent cell to daughter cell during cell division or from one generation to the next. The collection of all epigenetic changes in a genome is called an epigenome.

Epistasis

Epistasis is a circumstance where the expression of one gene is modified (e.g., masked, inhibited or suppressed) by the expression of one or more other genes.

Eugenics

Eugenics is a discredited belief that selective breeding for certain inherited human traits can improve the “fitness” of future generations. For eugenicists, “fitness” corresponded to a narrow view of humanity and society that developed directly from the ideologies and practices of scientific racism, colonialism, ableism and imperialism.

Evolution

Evolution, as related to genomics, refers to the process by which living organisms change over time through changes in the genome. Such evolutionary changes result from mutations that produce genomic variation, giving rise to individuals whose biological functions or physical traits are altered. Those individuals who are best at adapting to their surroundings leave behind more offspring than less well-adapted individuals. Thus, over successive generations (in some cases spanning millions of years), one species may evolve to take on divergent functions or physical characteristics or may even evolve into a different species.

Exome
An exome is the sequence of all the exons in a genome, reflecting the protein-coding portion of a genome. In humans, the exome is about 1.5% of the genome.

Exon
An exon is a region of the genome that ends up within an mRNA molecule. Some exons are coding, in that they contain information for making a protein, whereas others are non-coding. Genes in the genome consist of exons and introns.

-F-

Family History
A family history, as related to medicine, is a record of the diseases and health conditions of an individual and that person's biological family members, both living and deceased. A family history can help determine whether someone has an increased genetic risk of having or developing certain diseases, disorders or conditions. It is often recorded by drawing a pedigree (a family tree) that illustrates the relationships among individuals.

Fibroblast
A fibroblast is a type of cell that contributes to the formation of connective tissue, a fibrous cellular material that supports and connects other tissues or organs in the body. Fibroblasts secrete collagen proteins that help maintain the structural framework of tissues. They also play an important role in healing wounds. Obtained from a person through a simple skin biopsy, fibroblasts can be grown in the laboratory for use in genetic and other scientific studies of that individual.

First-Degree Relative
A first-degree relative is a family member who shares about half of their genetic information with specific other individuals in their family. First-degree relatives include an individual's parents, siblings and offspring.

Fluorescence In Situ Hybridization (FISH)
Fluorescence in situ hybridization (abbreviated FISH) is a laboratory technique used to detect and locate a specific DNA sequence on a chromosome. In this technique, the full set of chromosomes from an individual is affixed to a glass slide and then exposed to a "probe"—a small piece of purified DNA tagged with a fluorescent dye. The fluorescently labeled probe finds and then binds to its matching sequence within the set of chromosomes. With the use of a special microscope, the chromosome and sub-chromosomal location where the fluorescent probe bound can be seen.

Founder Effect
A founder effect, as related to genetics, refers to the reduction in genomic variability that occurs when a small group of individuals becomes separated from a larger population. Over time, the resulting new subpopulation will have genotypes and physical traits resembling the initial small, separated group, and these may be very different from the original larger population. A founder effect can also explain why certain inherited diseases are found more frequently in some limited population groups. In some cases, a founder effect can play a role in the emergence of new species.

Fragile X Syndrome
Fragile X syndrome is a genetic condition that affects a person's development, in particular their ability to learn and their social behavior. The syndrome results from mutations in a gene on the X chromosome. Because males have only one copy of the X chromosome, they are more likely to show severe symptoms if they inherit the mutated gene compared to females (who have two copies of the X chromosome).

Frameshift Mutation
A frameshift mutation in a gene refers to the insertion or deletion of nucleotide bases in numbers that are not multiples of three. This is important because a cell reads a gene's code in groups of three bases when making a protein. Each of these "triplet codons" corresponds to one of 20 different amino acids used to build a protein. If a mutation disrupts this normal reading frame, then the entire gene sequence following the mutation will be incorrectly read. This can result in the addition of the wrong amino acids to the protein and/or the creation of a codon that stops the protein from growing longer.

Fraternal Twins
Fraternal twins (also called dizygotic twins) result from the fertilization of two separate eggs with two different sperm during the same pregnancy. Fraternal twins may not have the same sex or appearance. They share half their genomes, just like any other siblings. In contrast, identical twins (or monozygotic twins) result from the fertilization of a single egg by a single sperm, with the fertilized egg then splitting into two. As a result, identical twins share the same genomes and are always the same sex.

-G-

Gamete
A gamete is a reproductive cell of an animal or plant. In animals, female gametes are called ova or egg cells, and male gametes are called sperm. Ova and sperm are haploid cells, with each cell carrying only one copy of each chromosome. During fertilization, a sperm and ovum unite to form a new diploid organism.

Gender
Gender is an evolving term, and there are different definitions. In a general sense, gender involves how a person identifies along a broad, fluid spectrum, rather than strictly by the sex assignment they received at birth. Sex refers to the physical differences (e.g., chromosome composition, anatomy, and physiology) among people who are born male, female, or intersex. The term gender may also be used to refer to the social or cultural constructs of “roles” or “norms” typically associated with being masculine or feminine. Gender does not always directly relate to sex assigned at birth.

Gene
The gene is considered the basic unit of inheritance. Genes are passed from parents to offspring and contain the information needed to specify physical and biological traits. Most genes code for specific proteins, or segments of proteins, which have differing functions within the body. Humans have approximately 20,000 protein-coding genes.

Gene Amplification
Gene amplification refers to an increase in the number of copies of a gene in a genome. Cancer cells, for example, sometimes produce multiple copies of a gene(s) in response to signals from other cells or the environment.

Gene–Environment Interaction
Gene–environment interaction refers to the interplay of genes (and, more broadly, genome function) and the physical and social environment. These interactions influence the expression of phenotypes. For example, most human traits and diseases are influenced by how one or more genes interact in complex ways with environmental factors, such as chemicals in the air or water, nutrition, ultraviolet radiation from the sun, and social context.

Gene Expression
Gene expression is the process by which the information encoded in a gene is used to either make RNA molecules that code for proteins or to make non-coding RNA molecules that serve other functions. Gene expression acts as an “on/off switch” to control when and where RNA molecules and proteins are made and as a “volume control” to determine how much of those products are made. The process of gene expression is carefully regulated, changing substantially under different conditions. The RNA and protein products of many genes serve to regulate the expression of other genes.

Gene Mapping
Gene mapping refers to the process of determining the location of genes on chromosomes. Today, the most efficient approach for gene mapping involves sequencing a genome and then using computer programs to analyze the sequence to identify the location of genes.

Gene Pool
A gene pool refers to the combination of all the genes (including alleles) present in a reproducing population or species. A large gene pool has extensive genomic diversity and is better able to withstand environmental challenges. Inbreeding contributes to a smaller gene pool, making populations or species less able to adapt and survive when faced with environmental challenges.

Gene Regulation
Gene regulation is the process used to control the timing, location and amount in which genes are expressed. The process can be complicated and is carried out by a variety of mechanisms, including through regulatory proteins and chemical modification of DNA. Gene regulation is key to the ability of an organism to respond to environmental changes.

Gene Therapy

Gene therapy is a technique that uses a gene(s) to treat, prevent or cure a disease or medical disorder. Often, gene therapy works by adding new copies of a gene that is broken, or by replacing a defective or missing gene in a patient's cells with a healthy version of that gene. Both inherited genetic diseases (e.g., hemophilia and sickle cell disease) and acquired disorders (e.g., leukemia) have been treated with gene therapy.

Genetic Ancestry

Genetic ancestry refers to information about the people that an individual is biologically descended from, including their genetic relationships. Genetic information can be combined with historical information to infer where an individual's distant ancestors lived.

Genetic Code

Genetic code refers to the instructions contained in a gene that tell a cell how to make a specific protein. Each gene's code uses the four nucleotide bases of DNA — adenine (A), cytosine (C), guanine (G) and thymine (T) — in various ways to spell out three-letter "codons" that specify which amino acid is needed at each position within a protein.

Genetic Counseling

Genetic counseling refers to guidance relating to genetic disorders that a specialized healthcare professional (genetic counselor) provides to an individual or family. A genetic counselor might provide information about how a genetic condition could affect an individual or family and/or interpret genetic tests designed to help estimate the risk of a disease. The genetic counselor conveys information to address the concerns of the individual or family, helps them make an informed decision about their medical situation and provides psychological counseling to help them adapt to their condition or risk.

Genetic Discrimination

Genetic discrimination refers to the unequal treatment of individuals based on an aspect of their genetic code (or genome), such as the risk for a genetic disorder. Genetic discrimination can involve such genomic information being used against individuals in a variety of circumstances, such as employment, health or disability insurance, or education. In a healthcare context, it can refer to people being treated in a discriminatory way based on their genetic status rather than by more relevant criteria.

Genetic Drift

Genetic drift is a mechanism of evolution characterized by random fluctuations in the frequency of a particular version of a gene (allele) in a population. Though it primarily affects small, isolated populations, the effects of genetic drift can be strong, sometimes causing traits to become overwhelmingly frequent or to disappear from a population.

Genetic Engineering

Genetic engineering (also called genetic modification) is a process that uses laboratory-based technologies to alter the DNA makeup of an organism. This may involve changing a single base pair (A-T or C-G), deleting a region of DNA or adding a new segment of DNA. For example, genetic engineering may involve adding a gene from one species to an organism from a different species to produce a desired trait. Used in research and industry, genetic engineering has been applied to the production of cancer therapies, brewing yeasts, genetically modified plants and livestock, and more.

Genetic Epidemiology

Genetic epidemiology is a field of science focused on the study of how genetic factors influence human traits, such as human health and disease. In many cases, the interaction of genes with the environment is also measured. Genetic epidemiologists seek to understand the causes, distribution and control of inherited disease in groups and the multifactorial causes of genetic diseases in populations.

Genetic Imprinting

Genetic imprinting (also called genomic imprinting) is the process by which only one copy of a gene in an individual (either from their mother or their father) is expressed, while the other copy is suppressed. Unlike genomic mutations that can affect the ability of inherited genes to be expressed, genetic imprinting does not affect the DNA sequence itself. Instead, gene expression is silenced by the epigenetic addition of chemical tags to the DNA during egg or sperm formation. Epigenetic tags on imprinted genes usually stay in place for the life of the individual.

Genetic Information Nondiscrimination Act (GINA)

The Genetic Information Nondiscrimination Act (abbreviated GINA) is federal legislation in the United States that protects individuals against discrimination based on their personal genetic information, as it applies to health insurance and employment. These protections are intended to encourage Americans to take advantage of genetic testing as part of their medical care. GINA was signed into law on May 22, 2008.

Genetic Map

A genetic map (also called a linkage map) shows the relative location of genetic markers (reflecting sites of genomic variants) on a chromosome. A genetic map is based on the concept of genetic linkage: the closer two markers are to each other on a chromosome, the greater the probability that they will be inherited together. By studying inheritance patterns, the relative order and location of genetic markers along a chromosome can be established.

Genetic Testing

Genetic testing is the use of a laboratory test to examine an individual's DNA for variations, typically performed in the context of medical care, ancestry studies or forensics. In a medical setting, the results of a genetic test can be used to confirm or rule out a suspected genetic disease. Results may also be used to determine the likelihood of parents passing on a genetic mutation to their offspring. Genetic testing may be performed prenatally or after birth. Genetic testing is also used to study the genomes of tumors in cancer cases.

Genetics

Genetics is the branch of biology concerned with the study of inheritance, including the interplay of genes, DNA variation and their interactions with environmental factors.

Genome

The genome is the entire set of DNA instructions found in a cell. In humans, the genome consists of 23 pairs of chromosomes located in the cell's nucleus, as well as a small chromosome in the cell's mitochondria. A genome contains all the information needed for an individual to develop and function.

Genome-Wide Association Study (GWAS)

A genome-wide association study (abbreviated GWAS) is a research approach used to identify genomic variants that are statistically associated with a risk for a disease or a particular trait. The method involves surveying the genomes of many people, looking for genomic variants that occur more frequently in those with a specific disease or trait compared to those without the disease or trait. Once such genomic variants are identified, they are typically used to search for nearby variants that contribute directly to the disease or trait.

Genomic Medicine

Genomic medicine is a medical discipline that involves using a person's genomic information as part of their clinical care. Other similar terms include individualized medicine, personalized medicine and precision medicine. For some conditions, genomic information can be used to help diagnose disease, predict outcomes and guide treatment.

Genomic Variation

Genomic variation refers to DNA sequence differences among individuals or populations. Some variants influence biological function (such as a mutation that causes a genetic disease), while others have no biological effects.

Genomics

Genomics is a field of biology focused on studying all the DNA of an organism — that is, its genome. Such work includes identifying and characterizing all the genes and functional elements in an organism's genome as well as how they interact.

Genotype

A genotype is a scoring of the type of variant present at a given location (i.e., a locus) in the genome. It can be represented by symbols. For example, BB, Bb, bb could be used to represent a given variant in a gene. Genotypes can also be represented by the actual DNA sequence at a specific location, such as CC, CT, TT. DNA sequencing and other methods can be used to determine the genotypes at millions of locations in a genome in a single experiment. Some genotypes contribute to an individual's observable traits, called the phenotype.

Germ Line

Germ line refers to the sex cells (eggs and sperm) that sexually reproducing organisms use to pass on their genomes from one generation to the next (parents to offspring). Egg and sperm cells are called germ cells, in contrast to the other cells of the body, which are called somatic cells.

Gigabase (Gb)

A gigabase (abbreviated Gb) is a unit of measurement used to help designate the length of DNA. One gigabase is equal to 1 billion bases.

GMO (Genetically Modified Organism)

GMO (short for “genetically modified organism”) is a plant, animal or microbe in which one or more changes have been made to the genome, typically using high-tech genetic engineering, in an attempt to alter the characteristics of an organism. Genes can be introduced, enhanced or deleted within a species, across species or even across kingdoms. GMOs may be used for a variety of purposes, such as making human insulin, producing fermented beverages and developing pesticide resistance in crop plants.

Guanine

Guanine (G) is one of the four nucleotide bases in DNA, with the other three being adenine (A), cytosine (C) and thymine (T). Within a double-stranded DNA molecule, guanine bases on one strand pair with cytosine bases on the opposite strand. The sequence of the four nucleotide bases encodes DNA’s information.

-H-

Haploid

Haploid refers to the presence of a single set of chromosomes in an organism’s cells. Sexually reproducing organisms are diploid (having two sets of chromosomes, one from each parent). In humans, only the egg and sperm cells are haploid.

Haplotype

A haplotype is a physical grouping of genomic variants (or polymorphisms) that tend to be inherited together. A specific haplotype typically reflects a unique combination of variants that reside near each other on a chromosome.

Heterozygous

Heterozygous, as related to genetics, refers to having inherited different versions (alleles) of a genomic marker from each biological parent. Thus, an individual who is heterozygous for a genomic marker has two different versions of that marker. By contrast, an individual who is homozygous for a marker has identical versions of that marker.

Histone

A histone is a protein that provides structural support for a chromosome. Each chromosome contains a long molecule of DNA, which must fit into the cell nucleus. To do that, the DNA wraps around complexes of histone proteins, giving the chromosome a more compact shape. Histones also play a role in the regulation of gene expression.

Homologous Recombination

Homologous recombination is a type of genetic recombination in which nucleotide sequences are exchanged between two similar or identical molecules of DNA. During the formation of egg and sperm cells (meiosis), paired chromosomes from the male and female parents align so that similar DNA sequences can cross over, or be exchanged, from one chromosome to the other. This exchanging of DNA is an important source of the genomic variation seen among offspring.

Homozygous

Homozygous, as related to genetics, refers to having inherited the same versions (alleles) of a genomic marker from each biological parent. Thus, an individual who is homozygous for a genomic marker has two identical versions of that marker. By contrast, an individual who is heterozygous for a marker has two different versions of that marker.

Human Genome Project

The Human Genome Project was a large international, collaborative effort that mapped and sequenced the human genome for the first time. Conducted from 1990 to 2003, the project was historic in its scope and scale as well as its groundbreaking approach for the free release of genomic data well ahead of publication, leading to a new ethos for data sharing in biomedical research.

Human Genome Reference Sequence

A human genome reference sequence is an accepted representation of the human genome sequence that is used by researchers as a standard for comparison to DNA sequences generated in their studies. The scientists responsible for assembling and updating such reference sequences aim to provide the highest-quality, best possible consensus representations of the sequence and structural diversity found in the human genome among populations. The genome reference sequence provides a general framework and is not the DNA sequence of a single person.

Huntington’s Disease

Huntington’s disease is a rare inherited disorder associated with the progressive loss of brain and muscle function. Symptoms usually develop during middle age and may include uncontrolled movements, loss of intellectual abilities and various emotional and psychiatric symptoms. Huntington’s disease is inherited as an autosomal dominant trait, meaning that a single mutated copy of the responsible gene (called *HTT*) is sufficient to cause the disease.

Hybridization

Hybridization, as related to genomics, is the process in which two complementary single-stranded DNA and/or RNA molecules bond together to form a double-stranded molecule. The bonding is dependent on the appropriate base-pairing across the two single-stranded molecules. Hybridization is an important process in various research and clinical laboratory techniques.

-I-

Identical Twins

Identical twins (also called monozygotic twins) result from the fertilization of a single egg by a single sperm, with the fertilized egg then splitting into two. Identical twins share the same genomes and are always of the same sex. In contrast, fraternal (dizygotic) twins result from the fertilization of two separate eggs with two different sperm during the same pregnancy. They share half of their genomes, just like any other siblings. Fraternal twins may not be of the same sex or have similar appearances.

In Situ Hybridization

In situ hybridization is a laboratory technique used to localize a sequence of DNA or RNA in a biological sample. In this technique, a biological sample consisting of tissue sections, cells or chromosomes from an individual is affixed to a glass slide and then exposed to a “probe”—a small piece of single-stranded DNA tagged with a chemical or fluorescent dye. The labeled probe finds and then binds to its matching sequence within the biological sample. The location of the bound probe can then be seen with the use of a microscope.

Inherited

Inherited, as related to genetics, refers to a trait or variants encoded in DNA and passed from parent to offspring during reproduction. Inheritance is determined by the rules of Mendelian genetics.

Insertion

An insertion, as related to genomics, is a type of mutation that involves the addition of one or more nucleotides into a segment of DNA. An insertion can involve the addition of any number of nucleotides, from a single nucleotide to an entire piece of a chromosome.

Intergenic Regions

Intergenic regions are the stretches of DNA located between genes. In humans, intergenic regions are non-protein-coding and comprise a large majority of the genome. Some intergenic DNA is known to regulate the expression of nearby genes.

Intron

An intron is a region that resides within a gene but does not remain in the final mature mRNA molecule following transcription of that gene and does not code for amino acids that make up the protein encoded by that gene. Most protein-coding genes in the human genome consist of exons and introns.

Inversion

An inversion in a chromosome occurs when a segment breaks off and reattaches within the same chromosome, but in reverse orientation. DNA may or may not be lost in the process.

-K-

Karyotype

A karyotype is an individual's complete set of chromosomes. The term also refers to a laboratory-produced image of a person's chromosomes isolated from an individual cell and arranged in numerical order. A karyotype may be used to look for abnormalities in chromosome number or structure.

Kilobase (kb)

A kilobase (abbreviated kb) is a unit of measurement used to help designate the length of DNA or RNA. One kilobase is equal to 1,000 bases.

Knockout

A knockout, as related to genomics, refers to the use of genetic engineering to inactivate or remove one or more specific genes from an organism. Scientists create knockout organisms to study the impact of removing a gene from an organism, which often allows them to then learn something about that gene's function.

-L-

Linkage

Linkage, as related to genetics and genomics, refers to the closeness of genes or other DNA sequences to one another on the same chromosome. The closer two genes or sequences are to each other on a chromosome, the greater the probability that they will be inherited together.

Locus

A locus, as related to genomics, is a physical site or location within a genome (such as a gene or another DNA segment of interest), somewhat like a street address. The plural of locus is loci.

LOD Score

An LOD (short for "logarithm of the odds") score is a statistical estimate of the relative probability that two loci (e.g., a disease-associated gene and another sequence of interest, such as a variant or another gene) are located near each other on a chromosome and are therefore likely to be inherited together.

LOH (Loss of Heterozygosity)

LOH (short for "loss of heterozygosity") refers to a type of mutation that results in the loss of one copy of a segment of DNA (typically containing a gene or group of genes). For most parts of the genome, human cells have two copies of any genomic segment — one from each parent — so in the case of LOH, only one copy would still be present.

Lyonization

Lyonization (also called X-inactivation) refers to the normal phenomenon in which one of the two X chromosomes in every cell of a female individual is inactivated during embryonic development. This inactivation prevents females from having twice as many X chromosome gene products as males, who possess only a single copy of the X chromosome. Lyonization is named after Mary F. Lyon, the British geneticist who discovered the phenomenon.

-M-

Mapping
Mapping refers to the process of determining the relative locations of landmarks or markers (such as genes, variants and other DNA sequences of interest) within a chromosome or genome. Historically, there have been two approaches for mapping: physical mapping, which established maps based on physical distances between landmarks, and genetic mapping, which established maps based on the frequency with which two landmarks are inherited together. Today, the most efficient approach for mapping involves sequencing a genome and then using computer programs to analyze the sequence to identify the locations of landmarks.

Marker
A marker (largely synonymous with the word "landmark" and often referred to as a genomic marker or a genetic marker) is a DNA sequence, typically with a known location in a genome. Markers can reflect random sequences, genomic variants or genes. Markers are used as signposts (or landmarks) in the construction of DNA and genome maps. Markers can also be used to track inheritance of traits or disease risk in families.

Megabase (Mb)
A megabase (abbreviated Mb) is a unit of measurement used to help designate the length of DNA. One megabase is equal to 1 million bases.

Meiosis
Meiosis is a type of cell division in sexually reproducing organisms that reduces the number of chromosomes in gametes (the sex cells, or egg and sperm). In humans, body (or somatic) cells are diploid, containing two sets of chromosomes (one from each parent). To maintain this state, the egg and sperm that unite during fertilization must be haploid, with a single set of chromosomes. During meiosis, each diploid cell undergoes two rounds of division to yield four haploid daughter cells — the gametes.

Mendel, Johann (Gregor)
Gregor Mendel was an Austrian monk in the 19th century who worked out the basic laws of inheritance through experiments with pea plants. In his monastery garden, Mendel performed thousands of crosses with pea plants, discovering how characteristics are passed down from one generation to the next — namely, dominant and recessive traits. Mendel's early experiments provided the basis of modern genetics.

Mendelian Inheritance
Mendelian inheritance refers to certain patterns of how traits are passed from parents to offspring. These general patterns were established by the Austrian monk Gregor Mendel, who performed thousands of experiments with pea plants in the 19th century. Mendel's discoveries of how traits (such as color and shape) are passed down from one generation to the next introduced the concept of dominant and recessive modes of inheritance.

Messenger RNA (mRNA)
Messenger RNA (abbreviated mRNA) is a type of single-stranded RNA involved in protein synthesis. mRNA is made from a DNA template during the process of transcription. The role of mRNA is to carry protein information from the DNA in a cell's nucleus to the cell's cytoplasm (watery interior), where the protein-making machinery reads the mRNA sequence and translates each three-base codon into its corresponding amino acid in a growing protein chain.

Metagenomics
Metagenomics is the study of the structure and function of entire nucleotide sequences isolated and analyzed from all the organisms (typically microbes) in a bulk sample. Metagenomics is often used to study a specific community of microorganisms, such as those residing on human skin, in the soil or in a water sample.

Metaphase
Metaphase is a stage during the process of cell division (mitosis or meiosis). Normally, individual chromosomes are spread out in the cell nucleus. During metaphase, the nucleus dissolves and the cell's chromosomes condense and move together, aligning in the center of the dividing cell. At this stage, the chromosomes are distinguishable when viewed through a microscope. Metaphase chromosomes are used in karyotyping, a laboratory technique for identifying chromosomal abnormalities.

Methylation

Methylation is a chemical modification of DNA and other molecules that may be retained as cells divide to make more cells. When found in DNA, methylation can alter gene expression. In this process, chemical tags called methyl groups attach to a particular location within DNA where they turn a gene on or off, thereby regulating the production of proteins that the gene encodes.

Microarray Technology

Microarray technology is a general laboratory approach that involves binding an array of thousands to millions of known nucleic acid fragments to a solid surface, referred to as a “chip.” The chip is then bathed with DNA or RNA isolated from a study sample (such as cells or tissue). Complementary base pairing between the sample and the chip-immobilized fragments produces light through fluorescence that can be detected using a specialized machine. Microarray technology can be used for a variety of purposes in research and clinical studies, such as measuring gene expression and detecting specific DNA sequences (e.g., single nucleotide polymorphisms, or SNPs).

Microbiome

The microbiome is the community of microorganisms (such as fungi, bacteria and viruses) that exists in a particular environment. In humans, the term is often used to describe the microorganisms that live in or on a particular part of the body, such as the skin or gastrointestinal tract. These groups of microorganisms are dynamic and change in response to a host of environmental factors, such as exercise, diet, medication and other exposures.

Microsatellite

Microsatellite, as related to genomics, refers to a short segment of DNA, usually one to six or more base pairs in length, that is repeated multiple times in succession at a particular genomic location. These DNA sequences are typically non-coding. The number of repeated segments within a microsatellite sequence often varies among people, which makes them useful as polymorphic markers for studying inheritance patterns in families or for creating a DNA fingerprint from crime scene samples.

Missense Mutation

A missense mutation is a DNA change that results in different amino acids being encoded at a particular position in the resulting protein. Some missense mutations alter the function of the resulting protein.

Mitochondrial DNA

Mitochondrial DNA is the circular chromosome found inside the cellular organelles called mitochondria. Located in the cytoplasm, mitochondria are the site of the cell’s energy production and other metabolic functions. Offspring inherit mitochondria — and as a result mitochondrial DNA — from their mother.

Mitosis

Mitosis is the process by which a cell replicates its chromosomes and then segregates them, producing two identical nuclei in preparation for cell division. Mitosis is generally followed by equal division of the cell’s content into two daughter cells that have identical genomes.

Monosomy

Monosomy refers to the condition in which only one chromosome from a pair is present in cells rather than the two copies usually found in diploid cells. When cells have one chromosome from a pair plus a portion of the second chromosome, this is referred to as partial monosomy. Monosomy, or partial monosomy, causes certain human diseases such as Turner syndrome and Cri du chat syndrome.

Mosaicism

Mosaicism refers to the presence of cells in a person that have a different genome from the body’s other cells. This difference could be due to a specific genomic variant, for example, or the addition or loss of a chromosome. The condition can stem from a genetic error that occurs after fertilization of an egg, during very early embryo development, or it could occur later in development. Mosaicism can affect any type of cell and does not always cause disease.

Mutagen

A mutagen is a chemical or physical agent capable of inducing changes in DNA called mutations. Examples of mutagens include tobacco products, radioactive substances, x-rays, ultraviolet radiation and a wide variety of chemicals. Exposure to a mutagen can produce DNA mutations that cause or contribute to certain diseases.

Mutation

A mutation is a change in the DNA sequence of an organism. Mutations can result from errors in DNA replication during cell division, exposure to mutagens or a viral infection. Germline mutations (that occur in eggs and sperm) can be passed on to offspring, while somatic mutations (that occur in body cells) are not passed on.

-N-

Nanopore DNA Sequencing

Nanopore DNA sequencing is a laboratory technique for determining the exact sequence of nucleotides, or bases, in a DNA molecule. The sequence of the bases (often referred to by the first letters of their chemical names: A, T, C and G) encodes the biological information that cells use to develop and operate. Nanopore DNA sequencing involves reading the code of single DNA strands as they are threaded through extremely tiny pores (nanopores) embedded within a membrane. As the DNA moves through the pore, it creates signals that can be converted to read each base. This approach offers a low-cost, rapid process for studying long stretches of DNA.

Nanotechnology

Nanotechnology (often shortened to nanotech) is the understanding and use of matter on an atomic and molecular scale for industrial purposes. Manipulating matter at nanoscale — between approximately 1 and 100 nanometers — holds potential for novel applications in many fields, including genomics, engineering, computer science and medicine.

Newborn Genetic Screening

Newborn screening is a set of laboratory tests performed on newborn babies to detect a set of known genetic diseases. Typically, this testing is performed on a blood sample obtained from a heel prick when the baby is two or three days old. In the United States, newborn screening is mandatory for a defined set of genetic diseases, although the exact set differs from state to state. Newborn screening tests focus on conditions for which early diagnosis is important to treating or preventing disease.

Next-Generation DNA Sequencing

DNA sequencing establishes the order of the bases that make up DNA. Next-generation DNA sequencing (abbreviated NGS) refers to the use of technologies for sequencing DNA that became available shortly after the completion of the Human Genome Project (which relied on the first-generation method of Sanger sequencing). Faster and cheaper than their predecessors, NGS technologies can sequence an entire human genome in a single day and for less than $1,000.

Non-Coding DNA

Non-coding DNA corresponds to the portions of an organism's genome that do not code for amino acids, the building blocks of proteins. Some non-coding DNA sequences are known to serve functional roles, such as in the regulation of gene expression, while other areas of non-coding DNA have no known function.

Nonsense Mutation

A nonsense mutation occurs in DNA when a sequence change gives rise to a stop codon rather than a codon specifying an amino acid. The presence of the new stop codon results in the production of a shortened protein that is likely non-functional.

Northern Blot

Northern blot is a laboratory analysis method used to study RNA. Specifically, purified RNA fragments from a biological sample (such as blood or tissue) are separated by using an electric current to move them through a sieve-like gel or matrix, which allows smaller fragments to move faster than larger fragments. The RNA fragments are transferred out of the gel or matrix onto a solid membrane, which is then exposed to a DNA probe labeled with a radioactive, fluorescent or chemical tag. The tag allows any RNA fragments containing complementary sequences with the DNA probe sequence to be visualized within the Northern blot.

Nuclear Membrane

The nuclear membrane is a double layer that encloses the cell's nucleus, where the chromosomes reside. The nuclear membrane serves to separate the chromosomes from the cell's cytoplasm and other contents. An array of small holes or pores in the nuclear membrane permits the selective passage of certain materials, such as nucleic acids and proteins, between the nucleus and cytoplasm.

Nucleic Acids
Nucleic acids are large biomolecules that play essential roles in all cells and viruses. A major function of nucleic acids involves the storage and expression of genomic information. Deoxyribonucleic acid, or DNA, encodes the information cells need to make proteins. A related type of nucleic acid, called ribonucleic acid (RNA), comes in different molecular forms that play multiple cellular roles, including protein synthesis.

Nucleolus
The nucleolus is a spherical structure found in the cell's nucleus whose primary function is to produce and assemble the cell's ribosomes. The nucleolus is also where ribosomal RNA genes are transcribed. Once assembled, ribosomes are transported to the cell cytoplasm, where they serve as the sites for protein synthesis.

Nucleopore
A nucleopore is one of a series of openings found in the cell's nuclear membrane. Nucleopores serve as channels for the selective transport of nucleic acids and proteins into and out of the cell nucleus.

Nucleosome
A nucleosome is the basic repeating subunit of chromatin packaged inside the cell's nucleus. In humans, about six feet of DNA must be packaged into a nucleus with a diameter less than a human hair, and nucleosomes play a key role in that process. A single nucleosome consists of about 150 base pairs of DNA sequence wrapped around a core of histone proteins. In forming a chromosome, the nucleosomes repeatedly fold in on themselves to tighten and condense the packaged DNA.

Nucleotide
A nucleotide is the basic building block of nucleic acids (RNA and DNA). A nucleotide consists of a sugar molecule (either ribose in RNA or deoxyribose in DNA) attached to a phosphate group and a nitrogen-containing base. The bases used in DNA are adenine (A), cytosine (C), guanine (G) and thymine (T). In RNA, the base uracil (U) takes the place of thymine. DNA and RNA molecules are polymers made up of long chains of nucleotides.

Nucleus
A nucleus, as related to genomics, is the membrane-enclosed organelle within a cell that contains the chromosomes. An array of holes, or pores, in the nuclear membrane allows for the selective passage of certain molecules (such as proteins and nucleic acids) into and out of the nucleus.

-O-

Oncogene
An oncogene is a mutated gene that has the potential to cause cancer. Before an oncogene becomes mutated, it is called a proto-oncogene, and it plays a role in regulating normal cell division. Cancer can arise when a proto-oncogene is mutated, changing it into an oncogene and causing the cell to divide and multiply uncontrollably. Some oncogenes work like an accelerator pedal in a car, pushing a cell to divide again and again. Others work like a faulty brake in a car parked on a hill, also causing the cell to divide unchecked.

Open Reading Frame
An open reading frame, as related to genomics, is a portion of a DNA sequence that does not include a stop codon (which functions as a stop signal). A codon is a DNA or RNA sequence of three nucleotides (a trinucleotide) that forms a unit of genomic information encoding a particular amino acid or signaling the termination of protein synthesis (stop codon). There are 64 different codons: 61 specify amino acids and 3 are used as stop codons. A long open reading frame is often part of a gene (that is, a sequence directly coding for a protein).

-P-

Pedigree

A pedigree, as related to genetics, is a chart that diagrams the inheritance of a trait or health condition through generations of a family. The pedigree particularly shows the relationships among family members and, when the information is available, indicates which individuals have a trait(s) of interest.

Peptide

A peptide is a short chain of amino acids (typically 2 to 50) linked by chemical bonds (called peptide bonds). A longer chain of linked amino acids (51 or more) is a polypeptide. The proteins manufactured inside cells are made from one or more polypeptides.

Pharmacogenomics

Pharmacogenomics (also called pharmacogenetics) is a component of genomic medicine that involves using a patient's genomic information to tailor the selection of drugs used in their medical management. In this way, pharmacogenomics aims to provide a more individualized (or precise) approach to the use of available medication in treating patients.

Phenotype

Phenotype refers to an individual's observable traits, such as height, eye color and blood type. A person's phenotype is determined by both their genomic makeup (genotype) and environmental factors.

Physical Map

A physical map, as related to genomics, is a graphical representation of physical locations of landmarks or markers (such as genes, variants and other DNA sequences of interest) within a chromosome or genome. A complete genome sequence is one type of physical map. Physical maps are used to identify genes or other sequences believed to play a role in health conditions or diseases. They are also valuable in providing an organizational framework for generating complete sequences of genomes.

Plasmid

A plasmid is a small circular DNA molecule found in bacteria and some other microscopic organisms. Plasmids are physically separate from chromosomal DNA and replicate independently. They typically have a small number of genes — notably, some associated with antibiotic resistance — and can be passed from one cell to another. Scientists use recombinant DNA methods to splice genes that they want to study into a plasmid. When the plasmid copies itself, it also makes copies of the inserted gene.

Point Mutation

A point mutation occurs in a genome when a single base pair is added, deleted or changed. While most point mutations are benign, they can also have various functional consequences, including changes in gene expression or alterations in encoded proteins.

Polydactyly

Polydactyly is a condition in which a person has more than the normal number of fingers or toes. It can occur in association with other physical anomalies or intellectual impairment, or it may occur as an isolated birth defect. Polydactyly can either be inherited or it can arise sporadically in an individual.

Polygenic Risk Score (PRS)

A polygenic risk score (abbreviated PRS) uses genomic information alone to assess a person's chances of having or developing a particular medical condition. A person's PRS is a statistical calculation based on the presence or absence of multiple genomic variants, without taking environmental or other factors into account.

Polygenic Trait

A polygenic trait is a characteristic, such as height or skin color, that is influenced by two or more genes. Because multiple genes are involved, polygenic traits do not follow the patterns of Mendelian inheritance. Many polygenic traits are also influenced by the environment and are called multifactorial.

Polymerase Chain Reaction (PCR)

Polymerase chain reaction (abbreviated PCR) is a laboratory technique for rapidly producing (amplifying) millions to billions of copies of a specific segment of DNA, which can then be studied in greater detail. PCR involves using short synthetic DNA fragments called primers to select a segment of the genome to be amplified, and then multiple rounds of DNA synthesis to amplify that segment.

Polymorphism

Polymorphism, as related to genomics, refers to the presence of two or more variant forms of a specific DNA sequence that can occur among different individuals or populations. The most common type of polymorphism involves variation at a single nucleotide (also called a single-nucleotide polymorphism, or SNP). Other polymorphisms can be much larger, involving longer stretches of DNA.

Population Genomics

Population genomics is the large-scale application of genomic technologies to study populations of individuals. For example, population genomics research can be used to study human ancestry, migrations and health.

Positional Cloning

Positional cloning is a laboratory approach used to locate the position of a disease-associated gene on a chromosome. Such a strategy can succeed even when nothing is known about the role of the gene's encoded protein in the disease. The technique typically relies on the use of known polymorphic markers whose inheritance can be traced through various members of families affected by the disease.

Precision Medicine

Precision medicine (generally considered analogous to personalized medicine or individualized medicine) is an innovative approach that uses information about an individual's genomic, environmental and lifestyle information to guide decisions related to their medical management. The goal of precision medicine is to provide more a precise approach for the prevention, diagnosis and treatment of disease.

Primer

A primer, as related to genomics, is a short single-stranded DNA fragment used in certain laboratory techniques, such as the polymerase chain reaction (PCR). In the PCR method, a pair of primers hybridizes with the sample DNA and defines the region that will be amplified, resulting in millions and millions of copies in a very short timeframe. Primers are also used in DNA sequencing and other experimental processes.

Proband

A proband is an individual who is affected by a genetic condition or who is concerned they are at risk. Usually, the proband is the first person in a family who brings the concern of a genetic disorder to the attention of healthcare professionals.

Promoter

A promoter, as related to genomics, is a region of DNA upstream of a gene where relevant proteins (such as RNA polymerase and transcription factors) bind to initiate transcription of that gene. The resulting transcription produces an RNA molecule (such as mRNA).

Protein

Proteins are large, complex molecules that play many important roles in the body. They are critical to most of the work done by cells and are required for the structure, function and regulation of the body's tissues and organs. A protein is made up of one or more long, folded chains of amino acids (each called a polypeptide), whose sequences are determined by the DNA sequence of the protein-encoding gene.

Pseudogene

A pseudogene is a segment of DNA that structurally resembles a gene but is not capable of coding for a protein. Pseudogenes are most often derived from genes that have lost their protein-coding ability due to accumulated mutations that have occurred over the course of evolution.

-R-

Race
Race is a social construct used to group people. Race was constructed as a hierarchal human-grouping system, generating racial classifications to identify, distinguish and marginalize some groups across nations, regions and the world. Race divides human populations into groups often based on physical appearance, social factors and cultural backgrounds.

Recessive Traits and Alleles
Recessive, as related to genetics, refers to the relationship between an observed trait and the two inherited versions of a gene related to that trait. Individuals inherit two versions of each gene, known as alleles, from each parent. In the case of a recessive trait, the alleles of the trait-causing gene are the same, and both (recessive) alleles must be present to express the trait. A recessive allele does not produce a trait at all when only one copy is present. This contrasts to a dominant trait, which requires that only one of the two alleles be present to express the trait.

Recombinant DNA Technology
Recombinant DNA technology involves using enzymes and various laboratory techniques to manipulate and isolate DNA segments of interest. This method can be used to combine (or splice) DNA from different species or to create genes with new functions. The resulting copies are often referred to as recombinant DNA. Such work typically involves propagating the recombinant DNA in a bacterial or yeast cell, whose cellular machinery copies the engineered DNA along with its own.

Repressor
A repressor, as related to genomics, is a protein that inhibits the expression of one or more genes. The repressor protein works by binding to the promoter region of the gene(s), which prevents the production of messenger RNA (mRNA). Repressor proteins are essential for the regulation of gene expression in cells.

Restriction Enzyme
A restriction enzyme is a protein isolated from bacteria that cleaves DNA sequences at sequence-specific sites, producing DNA fragments with a known sequence at each end. The use of restriction enzymes is critical to certain laboratory methods, including recombinant DNA technology and genetic engineering.

Restriction Fragment Length Polymorphism (RFLP)
Restriction fragment length polymorphism (abbreviated RFLP) refers to differences (or variations) among people in their DNA sequences at sites recognized by restriction enzymes. Such variation results in different sized (or length) DNA fragments produced by digesting the DNA with a restriction enzyme. RFLPs can be used as genetic markers, which are often used to follow the inheritance of DNA through families.

Retrovirus
A retrovirus is a virus that uses RNA as its genomic material. Upon infection with a retrovirus, a cell converts the retroviral RNA into DNA, which in turn is inserted into the DNA of the host cell. The cell then produces more retroviruses, which infect other cells. Many retroviruses are associated with diseases, including AIDS and some forms of cancer.

Ribonucleic Acid (RNA)
Ribonucleic acid (abbreviated RNA) is a nucleic acid present in all living cells that has structural similarities to DNA. Unlike DNA, however, RNA is most often single-stranded. An RNA molecule has a backbone made of alternating phosphate groups and the sugar ribose, rather than the deoxyribose found in DNA. Attached to each sugar is one of four bases: adenine (A), uracil (U), cytosine (C) or guanine (G). Different types of RNA exist in cells: messenger RNA (mRNA), ribosomal RNA (rRNA) and transfer RNA (tRNA). In addition, some RNAs are involved in regulating gene expression. Certain viruses use RNA as their genomic material.

Ribosome
A ribosome is an intercellular structure made of both RNA and protein, and it is the site of protein synthesis in the cell. The ribosome reads the messenger RNA (mRNA) sequence and translates that genetic code into a specified string of amino acids, which grow into long chains that fold to form proteins.

Risk

Risk, as related to genetics, refers to the probability that an individual will be affected by a particular heritable or genetic disorder. Both a person's genome and environmental exposures can influence risk. An individual's risk may be higher because they inherit a genetic variant (or allele) in one gene or a combination of many variants in different genes that increases susceptibility to or overtly causes a disorder. Other individuals may be at higher risk because they have been exposed to one or more environmental factors that promote the development of a certain disorder.

-S-

Scientific Racism

Scientific racism is a historical pattern of ideologies that generate and perpetuate pseudoscientific racist beliefs that lead to racial bias and discrimination in science and research. Leading scientists across scientific institutions in the 19th and early 20th centuries were proponents of such ideologies. By the mid-20th century, pseudoscientific racist beliefs were widely disproven. However, evidence shows that scientific racism persists in science and research.

Secondary Genomic Finding

A secondary genomic finding refers to a genomic variant, found through the analysis of a person's genome, that is of potential medical value yet is unrelated to the initial reason for examining the person's genome. In certain cases, a secondary genomic finding might offer clinicians the chance to identify a previously unrecognized risk for disease that could change the medical management of that patient and potentially prevent or more effectively treat the disease.

Sex Chromosome

A sex chromosome is a type of chromosome involved in sex determination. Humans and most other mammals have two sex chromosomes, X and Y, that in combination determine the sex of an individual. Females have two X chromosomes in their cells, while males have one X and one Y.

Sex-Linked

Sex-linked, as related to genetics, refers to characteristics (or traits) that are influenced by genes carried on the sex chromosomes. In humans, the term often refers to traits or disorders influenced by genes on the X chromosome, as it contains many more genes than the smaller Y chromosome. Males, who have only a single copy of the X chromosome, are more likely to be affected by a sex-linked disorder than females, who have two copies. In females, the presence of a second, non-mutated copy may cause different, milder, or no symptoms of a sex-linked disorder.

Shotgun Sequencing

Shotgun sequencing is a laboratory technique for determining the DNA sequence of an organism's genome. The method involves randomly breaking up the genome into small DNA fragments that are sequenced individually. A computer program looks for overlaps in the DNA sequences, using them to reassemble the fragments in their correct order to reconstitute the genome.

Single Nucleotide Polymorphism (SNP)

A single nucleotide polymorphism (abbreviated SNP, pronounced snip) is a genomic variant at a single base position in the DNA. Scientists study if and how SNPs in a genome influence health, disease, drug response and other traits.

Somatic Cells

Somatic cells are the cells in the body other than sperm and egg cells (which are called germ cells). In humans, somatic cells are diploid, meaning they contain two sets of chromosomes, one inherited from each parent. DNA mutations in somatic cells can affect an individual, but they cannot be passed on to their offspring.

Southern Blot

Southern blot analysis is a laboratory method used to study DNA. Specifically, purified DNA from a biological sample (such as blood or tissue) is digested with a restriction enzyme(s), and the resulting DNA fragments are separated by using an electric current to move them through a sieve-like gel or matrix, which allows smaller fragments move faster than larger fragments. The DNA fragments are transferred out of the gel or matrix onto a solid membrane, which is then exposed to a DNA probe labeled with a radioactive, fluorescent or chemical tag. The tag allows any DNA fragments containing complementary sequences with the DNA probe sequence to be visualized within the Southern blot. The method is named for its creator, British molecular biologist Edwin Southern.

Stop Codon
A stop codon is a sequence of three nucleotides (a trinucleotide) in DNA or messenger RNA (mRNA) that signals a halt to protein synthesis in the cell. There are 64 different trinucleotide codons: 61 specify amino acids and 3 are stop codons (i.e., UAA, UAG and UGA).

Substitution
Substitution, as related to genomics, is a type of mutation in which one nucleotide is replaced by a different nucleotide. The term can also refer to the replacement of one amino acid in a protein with a different amino acid.

Susceptibility
Susceptibility, as related to genetics, refers to the state of being predisposed to, or sensitive to, developing a certain disease. An individual's disease susceptibility is influenced by a combination of genetic and environmental factors.

Syndrome
A syndrome, as related to genetics, is a group of traits or conditions that tend to occur together and characterize a recognizable disease. Some syndromes have a genetic cause.

-T-

Tandem Repeat
A tandem repeat is a sequence of two or more DNA bases that is repeated numerous times in a head-to-tail manner on a chromosome. Tandem repeats are generally present in non-coding DNA. In some cases, tandem repeats can serve as genetic markers to track inheritance in families. They can also be useful for DNA fingerprinting in forensic studies.

Telomere
A telomere is a region of repetitive DNA sequences at the end of a chromosome. Telomeres protect the ends of chromosomes from becoming frayed or tangled. Each time a cell divides, the telomeres become slightly shorter. Eventually, they become so short that the cell can no longer divide successfully, and the cell dies.

Thymine
Thymine (T) is one of the four nucleotide bases in DNA, with the other three being adenine (A), cytosine (C) and guanine (G). Within a double-stranded DNA molecule, thymine bases on one strand pair with adenine bases on the opposite strand. The sequence of the four nucleotide bases encodes DNA's information.

Trait
A trait, as related to genetics, is a specific characteristic of an individual. Traits can be determined by genes, environmental factors or by a combination of both. Traits can be qualitative (such as eye color) or quantitative (such as height or blood pressure). A given trait is part of an individual's overall phenotype.

Transcription
Transcription, as related to genomics, is the process of making an RNA copy of a gene's DNA sequence. This copy, called messenger RNA (mRNA), carries the gene's protein information encoded in DNA. In humans and other complex organisms, mRNA moves from the cell nucleus to the cell cytoplasm (watery interior), where it is used for synthesizing the encoded protein.

Transfer RNA (tRNA)
Transfer RNA (abbreviated tRNA) is a small RNA molecule that plays a key role in protein synthesis. Transfer RNA serves as a link (or adaptor) between the messenger RNA (mRNA) molecule and the growing chain of amino acids that make up a protein. Each time an amino acid is added to the chain, a specific tRNA pairs with its complementary sequence on the mRNA molecule, ensuring that the appropriate amino acid is inserted into the protein being synthesized.

Transgenic
Transgenic refers to an organism or cell whose genome has been altered by the introduction of one or more foreign DNA sequences from another species by artificial means. Transgenic organisms are generated in the laboratory for research purposes.

Translation
Translation, as related to genomics, is the process through which information encoded in messenger RNA (mRNA) directs the addition of amino acids during protein synthesis. Translation takes place on ribosomes in the cell cytoplasm, where mRNA is read and translated into the string of amino acid chains that make up the synthesized protein.

Translocation
A translocation, as related to genetics, occurs when a chromosome breaks and the (typically two) fragmented pieces re-attach to different chromosomes. The detection of chromosomal translocations can be important for the diagnosis of certain genetic diseases and disorders.

Tumor Suppressor Gene
A tumor suppressor gene encodes a protein that acts to regulate cell division, keeping it in check. When a tumor suppressor gene is inactivated by a mutation, the protein it encodes is not produced or does not function properly, and as a result, uncontrolled cell division may occur. Such mutations may contribute to the development of a cancer.

-U-

Uracil
Uracil (U) is one of the four nucleotide bases in RNA, with the other three being adenine (A), cytosine (C) and guanine (G). In RNA, uracil pairs with adenine. In a DNA molecule, the nucleotide thymine (T) is used in place of uracil.

-V-

Variant of Uncertain Significance (VUS)
When analysis of a patient’s genome identifies a variant, but it is unclear whether that variant is actually connected to a health condition, the finding is called a variant of uncertain significance (abbreviated VUS). In many cases, these variants are so rare in the population that little information is available about them. Typically, more information is required to determine if the variant is disease related. Such information may include more extensive population data, functional studies, and tracing the variant in other family members who have or do not have the same health condition.

Vector
A vector, as related to molecular biology, is a DNA molecule (often plasmid or virus) that is used as a vehicle to carry a particular DNA segment into a host cell as part of a cloning or recombinant DNA technique. The vector typically assists in replicating and/or expressing the inserted DNA sequence inside the host cell.

Virus
A virus is an infectious microbe consisting of a segment of nucleic acid (either DNA or RNA) surrounded by a protein coat. A virus cannot replicate alone; instead, it must infect cells and use components of the host cell to make copies of itself. Often, a virus ends up killing the host cell in the process, causing damage to the host organism. Well-known examples of viruses causing human disease include AIDS, COVID-19, measles and smallpox.

-X-

X Chromosome

The X chromosome is one of the two sex chromosomes that are involved in sex determination. Humans and most other mammals have two sex chromosomes (X and Y) that in combination determine the sex of an individual. Females have two X chromosomes in their cells, while males have one X and one Y.

X-Linked

X-linked, as related to genetics, refers to characteristics or traits that are influenced by genes on the X chromosome. Humans and most other mammals have two sex chromosomes, X and Y. Females have two X chromosomes in their cells, while males have one X and one Y. In the case of an X-linked disease, it is usually males that are affected because they have a single copy of the X chromosome that carries the disease-causing mutation. In females, the presence of a second, non-mutated copy may cause different, milder, or no symptoms of a sex-linked disorder.

-Y-

Y Chromosome

The Y chromosome is one of the two sex chromosomes that are involved in sex determination. Humans and most other mammals have two sex chromosomes (X and Y) that in combination determine the sex of an individual. Females have two X chromosomes in their cells, while males have one X and one Y.

National Human Genome
Research Institute

genome.gov/tg

The **Forefront**
of **Genomics**

April 2022

© American College of Medical Genetics and Genomics

REVIEW | Genetics inMedicine

Open

Implementing genomic medicine in the clinic: the future is here

Teri A. Manolio, MD, PhD[1], Rex L. Chisholm, PhD[2], Brad Ozenberger, PhD[1], Dan M. Roden, MD[3], Marc S. Williams, MD[4,5], Richard Wilson, PhD[6], David Bick, MD[7], Erwin P. Bottinger, MD[8], Murray H. Brilliant, PhD[9], Charis Eng, MD, PhD[10], Kelly A. Frazer, PhD[11], Bruce Korf, MD, PhD[12], David H. Ledbetter, PhD[5], James R. Lupski, MD, PhD[13], Clay Marsh, MD[14], David Mrazek, MD[15], Michael F. Murray, MD[16], Peter H. O'Donnell, MD[17], Daniel J. Rader, MD[18], Mary V. Relling, PharmD[19], Alan R. Shuldiner, MD[20], David Valle, MD[21], Richard Weinshilboum, MD[22], Eric D. Green, MD, PhD[1] and Geoffrey S. Ginsburg, MD, PhD[23]

Although the potential for genomics to contribute to clinical care has long been anticipated, the pace of defining the risks and benefits of incorporating genomic findings into medical practice has been relatively slow. Several institutions have recently begun genomic medicine programs, encountering many of the same obstacles and developing the same solutions, often independently. Recognizing that successful early experiences can inform subsequent efforts, the National Human Genome Research Institute brought together a number of these groups to describe their ongoing projects and challenges, identify common infrastructure and research needs, and outline an implementation framework for investigating and introducing similar programs elsewhere. Chief among the challenges were limited evidence and consensus on which genomic variants were medically relevant; lack of reimbursement for genomically driven interventions; and burden to patients and clinicians of assaying, reporting, intervening, and following up genomic findings. Key infrastructure needs included an openly accessible knowledge base capturing sequence variants and their phenotypic associations and a framework for defining and cataloging clinically actionable variants. Multiple institutions are actively engaged in using genomic information in clinical care. Much of this work is being done in isolation and would benefit from more structured collaboration and sharing of best practices.

Genet Med advance online publication 10 January 2013

Key Words: medical genomics; practice standards

INTRODUCTION

The potential for genomics to contribute to clinical care has long been recognized, and many optimistic scenarios for clinical use of information about a patient's genome have been proposed.[1–3] The pace of realizing this potential has appeared slow to some,[4,5] although clinical adoption of scientific discoveries has been estimated to take up to 17 years[6] and a recent genetic example[7] required 18 years after the initial 1991 report.[8] Indeed, relatively robust genotype–phenotype associations for common, complex diseases only began to become available around 2005.[9] Yet several academic medical centers and integrated health systems have already begun programs for implementing genomic medicine, which we define here as using an individual patient's genotypic information in his or her clinical care. This definition encompasses both Mendelian and multigenic complex diseases, and at present probably focuses most on single Mendelian variants of large effect, although emphasis is expected to shift soon to assaying and using multiple variants simultaneously in clinical care. This is not only because of an expanding knowledge base but also because the logistic complexities and costs of assessing genomic variation on a large scale are approaching those of one-at-a-time testing of individual variants, permitting a more holistic approach to incorporating genomic findings into clinical care.

These initial and separate forays have encountered similar obstacles and developed many of the same nascent solutions,

---

[1]National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland, USA; [2]Center for Genetic Medicine, Northwestern University, Chicago, Illinois, USA; [3]Departments of Medicine and Pharmacology, Vanderbilt University School of Medicine, Nashville, Tennessee, USA; [4]Clinical Genetics Institute, Intermountain Healthcare, Salt Lake City, Utah, USA; [5]Genomic Medicine Institute, Geisinger Health System, Danville, Pennsylvania, USA; [6]The Genome Institute, Washington University School of Medicine, St. Louis, Missouri, USA; [7]Division of Genetics, Department of Pediatrics, Medical College of Wisconsin, Milwaukee, Wisconsin, USA; [8]The Charles Bronfman Institute for Personalized Medicine, Mount Sinai School of Medicine, New York, New York, USA; [9]Center for Human Genetics, Marshfield Clinic, Marshfield, Wisconsin, USA; [10]Genomic Medicine Institute, Cleveland Clinic Lerner Research Institute, Cleveland, Ohio, USA; [11]Department of Pediatrics, Division of Genome Information Sciences, University of California San Diego, La Jolla, California, USA; [12]Department of Genetics, University of Alabama at Birmingham, Birmingham, Alabama, USA; [13]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas, USA; [14]Center for Personalized Health Care, Ohio State University Medical Center, Columbus, Ohio, USA; [15]Department of Psychiatry and Psychology, Mayo Clinic, Rochester, Minnesota, USA; [16]Genetics Division, Brigham and Women's Hospital, Boston, Massachusetts, USA; [17]Department of Medicine, University of Chicago Medical Center, Chicago, Illinois, USA; [18]Division of Translational Medicine and Human Genetics, Department of Medicine, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, Pennsylvania, USA; [19]Department of Pharmaceutical Sciences, St. Jude Children's Research Hospital, Memphis, Tennessee, USA; [20]Program in Personalized and Genomic Medicine, Department of Medicine, University of Maryland School of Medicine, Baltimore, Maryland, USA; [21]Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA; [22]Department of Molecular Pharmacology and Experimental Therapeutics, Mayo Clinic College of Medicine, Rochester, Minnesota, USA; [23]Genomic Medicine, Duke Institute for Genome Sciences & Policy; Center for Personalized Medicine, Duke University, Durham, North Carolina, USA. Correspondence: Teri A. Manolio (manolio@nih.gov)

Submitted 16 April 2012; accepted 30 October 2012; advance online publication 10 January 2013. doi:10.1038/gim.2012.157

often quite independently. In expectation that sharing of lessons learned in these efforts can facilitate broader and more effective implementation of genomic medicine, the National Human Genome Research Institute in collaboration with several leaders in this area (R.L.C., G.S.G., D.M.R., M.S.W., and R.Wilson) brought together about 20 groups working to implement genomic medicine for a Genomic Medicine Colloquium in June 2011. This article summarizes the projects described by these groups, describes challenges they have encountered in implementation, identifies common infrastructure and research needs, and outlines an implementation framework for introducing similar genomic medicine programs more broadly.

BRIEF LANDSCAPE OF GENOMIC MEDICINE PROJECTS

Participating sites reported a broad range of genomic medicine activities, both in pilot and full implementation forms (**Table 1**). These included genotyping (used here and throughout to refer to the detection of genetic variants by single-nucleotide polymorphism assay, sequencing, or other technologies) of somatic mutations in malignant tumors to guide treatment decisions. The expanding scope and growing acceptance of such techniques places tumor genotyping at the forefront of genotype-directed care. All the participating sites (and most cancer centers nationwide) conduct tumor genotyping of cancers such as melanoma and those of the breast, colon, and lung for the targeting of therapy. Several also use genotyping to select patients for interventional protocols, but the main focus here was on the use of genomics outside of cancer diagnosis and treatment.

Several centers have begun targeted screening for highly penetrant germline mutations for Lynch syndrome and in *BRCA1/2*[10–12] to identify genetically at-risk individuals. With appropriate immunohistochemical or other evidence suggestive of germline mutations, patients can be actively sought and counseled regarding the importance of genomic testing for their own treatment and for potential preventive care of their relatives. Direct outreach to patients at the Cleveland Clinic, for example, increased the uptake of germline testing in patients with colorectal cancer by nearly sixfold—from 14% with information provided directly to surgeons to 80% with a genetic counselor seeing patients at follow-up examinations (C.E., unpublished data). Highly penetrant variants such as these provide robust, evidence-based paradigms to help move a clinical site toward implementing broader genomic medicine approaches. More important, perhaps, successful implementation of screening programs at early adopter sites provides valuable lessons and refined approaches that may be adopted elsewhere.[13]

Table 1 Examples of genomic medicine projects currently in implementation at participating sites

Tumor-based genotype-driven treatment
Risk/susceptibility testing in relatives of patients with mutation-bearing cancer (Lynch syndrome, *BRCA1/2*, etc.)
Family history collection for assessment of individual risk
CYP2C19 and antiplatelet therapy
Other genotype-driven treatment decisions
Whole-genome/whole-exome sequencing for unknown disease diagnosis
Complex disease risk advice (myocardial infarction and type 2 diabetes)

Self-reported family history information can also be used in risk assessment for individual patients with documented clinical validity and utility,[14,15] despite its potential weaknesses such as incomplete or inaccurate reporting.[16] Family history is generally not difficult to collect, especially with electronic data collection tools such as the Surgeon General's My Family Health Portrait,[17] although integrating the information into an independent electronic medical record (EMR) is not trivial. Clinicians are generally familiar with family history and appreciate its importance, and evidence-based guidelines are available for several conditions.[12,14] A Duke University pilot project, for example, is collecting three-generation family history information on 48 diseases using a Web-based computerized tool.[18] Patient-entered information is integrated into the medical record and generates a pedigree, tabular family history, and reports for the patient and clinician, as well as decision support information for four conditions (breast, ovarian, and colorectal cancer and venous thrombosis). Outcomes being assessed include patient and clinician satisfaction and ease of use; changes in patient behaviors such as diet and smoking; changes in physician behaviors such as screening and referral; and estimates of sensitivity, specificity, net reclassification, and cost.[18] Similarly, a Cleveland Clinic prototype collects three-generation family histories for the assessment of all inherited cancer syndromes.[19] Intermountain Healthcare has also deployed a patient-entered family tool ("Our Family Health") in their electronic patient portal.[20]

Pharmacogenomics provides additional opportunities in genomic medicine, particularly if appropriate decision support tools present relevant data to clinicians only when needed. Patients genetically unable to activate prodrugs such as clopidogrel, tamoxifen, and codeine, for example, do not need their genetic data presented to their clinicians unless and until these drugs are about to be prescribed. Although "reactive" genotyping for pharmacogenomic variants can be ordered at many sites, turnaround is often slow and uptake low. To address this, the Vanderbilt University Medical Center and St. Jude Children's Research Hospital have both implemented programs to conduct preemptive genotyping in patients very likely to receive medications with relevant genetically based dosing algorithms. At Vanderbilt, for example, selected patients, including those undergoing coronary angiography for acute coronary syndromes receive genotyping on a 184-variant platform that includes *CYP2C19* alleles before being prescribed clopidogrel,[21] based on a review of available evidence of poorer outcomes in persons with variant *CYP2C19* alleles.[22–24] At St. Jude Children's Research Hospital, array-based testing for 225 genes is performed, and results for those genes with the strongest clinical evidence are placed in the EMR.[25] Clinical pharmacogenomic testing is expanding to several sites affiliated with the Pharmacogenomics Research Network (PGRN)[26] with a goal of implementing proactive genotyping of a large number of variants, such as those identified by PharmGKB[27] or the US Food and Drug Administration (FDA).[28]

Preemptive testing, embedding this information into the EMR with appropriate decision support, and then seamlessly presenting it at the point of care when needed, rather than requiring tests to be ordered each time a relevant medication is prescribed, could help to minimize delay, cost, lack of follow-through, and duplicate testing. This is a testable hypothesis and one that is likely to be explored in programs such as the Electronic Medical Records and Genomics (eMERGE) Network.[29] Preemptive pharmacogenomic information could be even more useful were informed patients able to access it and work proactively with their clinicians in making treatment decisions. Putting patients in control of their genomic information is another potential strategy for genomic medicine implementation that needs evaluation. Such implementation strategies could be evaluated by measuring improvement, or lack thereof, in efficiency, cost, and health outcomes.

Whole-exome (protein-coding exons plus other highly conserved regions) and whole-genome sequencing are beginning to be used for identifying genetic causes of rare, unknown conditions, particularly those presenting in childhood or in strongly affected pedigrees for which novel mutations are more easily detected.[30] The recent cases of a young boy with severe, atypical, refractory inflammatory bowel disease[31] and of three families with severe arterial calcifications[32] demonstrate the potential for this approach to detect novel disease-causing mutations and point the way toward appropriate therapies. The Medical College of Wisconsin and Children's Hospital of Wisconsin have implemented a genome sequencing program to which clinicians can nominate patients who remain undiagnosed after appropriate clinical evaluation and testing. Nominations are reviewed by a multidisciplinary patient selection committee chaired by the hospital's chief medical officer. Patients and their families undergo 6–8 h of assessment and counseling before enrollment. Sequencing is provided by a commercial laboratory certified by the College of American Pathologists (CAP) Laboratory Accreditation Program and the Clinical Laboratory Improvement Amendments (CLIA) Program. The detected variants are analyzed using an in-house, validated software package, focusing first on specific candidate genes nominated by the referring clinicians to limit the chances of incidental, potentially unrelated findings. Analysis is then broadened if initially nominated genes are uninformative. Potentially relevant variants are confirmed by an independent method, and those results are placed in the medical record. The Baylor College of Medicine has implemented a Whole Genome Laboratory[33] with "in-house" sequencing and analysis of variants in a CLIA/CAP environment with review and sign-out by a board-certified clinical molecular geneticist. A number of other centers (Geisinger Clinic, Partners Healthcare, Washington University) and commercial laboratories such as Ambry Genetics,[34] GeneDx,[35] and SeqWright[36] are also now offering whole-genome sequencing that can be used in clinical care.

These four briefly described applications—tumor-based screening, family history–directed decision support, pharmacogenomics, and diagnostic genome sequencing—demonstrate that genomic medicine is no longer on the threshold; it has arrived. Although much of this work is being pursued as part of research programs, some has been adopted institutionally and offered in a CLIA-certified laboratory environment as part of regular clinical care.[13] Evaluating the impact of such programs and expanding their reach to diverse settings and populations is a high priority within the National Human Genome Research Institute's research agenda and mission. Information from such programs is helping to provide a needed evidence base to support more widespread implementation and reimbursement. Consideration of the barriers encountered and strategies used to surmount them by "early adopter" genomic medicine sites may help speed their evaluation and facilitate their expansion to the mainstream of clinical care.

CHALLENGES IN IMPLEMENTATION

Numerous challenges and barriers have been encountered in launching genomic medicine projects and many common solutions have emerged (**Table 2**). The greatest challenge may be the lack of appreciation by clinicians, health-care institutions, and payers of the potential for genomics to improve patient care, as compelling evidence of clinical validity and utility is limited at present.[37] Even for genomic applications with proven validity and utility, such as family history, there is lack of adoption and implementation, suggesting the lack of evidence is far from being the only barrier.

Cost concerns and institutional inertia typically demand convincing arguments and hard data before clinical practice is changed, yet the genome-wide nature of results to be expected by incorporating dense genotyping technologies into clinical care will soon force clinicians and institutions to deal with genomic information for which evidence may be quite limited. Evidentiary thresholds may thus need to be aligned with the intended use of the information, as the impact of genotype-driven care on morbidity and mortality cannot be tested for every variant.

Institutions have frequently relied on expert panels and local committees to evaluate available evidence and recommend particular new initiatives in genomic medicine, such as testing for specific pharmacogenomic variants or estimating risk to carriers. Such panels have tended to work in isolation from institution to institution, surveying the same evidence and often coming to similar conclusions. Far better would be to harness the collective knowledge of these groups in a more systematic way, distributing among them the genomic medicine issues for which evidence is compiled and evaluated, such as antidepressant therapy, coronary disease risk assessment, or risk in relatives of patients with colorectal cancer. These groups could then use a mutually accepted protocol for evaluation, with conclusions of each process shared among all. This is similar to the approach taken by the Centers for Disease Control and Prevention's Evaluation of Genomic Applications in Practice and Prevention (EGAPP),[38]

Table 2 Challenges in implementation and potential solutions

Challenge	Potential solutions
Limited evidence and conflicting interpretation of benefit/value	Generate evidence of clinical utility of genomic medicine interventions
	Tailor needs for evidence against potential benefits and risks
	Convene expert panels to develop guidelines based on best available evidence
	Engage nonaffiliated community-based practices, where institutional resistance may be minimal, to assist in developing evidence
	Determine process outcomes of incorporating genomic information into EMRs
	Widely publicize successes
	Organize consortia to conduct multicenter trials of genotype-directed therapy when appropriate
Lack of institutional and clinician acceptance	Establish institutional advisory committee(s) involving senior leadership to evaluate evidence, recommend, and monitor implementation
	Engage early adopters and clinical champions in demonstration efforts
	Conduct pilot projects in early adopter clinical environments to develop results sufficient for follow-on funding
	Obtain transinstitutional commitment at highest levels involving all relevant departments and stakeholders
	Utilize internal pilot funding to catalyze initiation
	Build clinician acceptance of genetics professionals by judiciously integrating genetic counselors and/or geneticists in nongenetics clinical services throughout primary institution and affiliates
	Bring fragmented expertise for advancing genomic medicine under one transinstitutional center or institute
	Harness institutional quality improvement processes to assess and demonstrate value
Limited access to genomic medicine expertise and testing	Use research screening assays on site and confirm clinically actionable findings with rapid, cost-effective CLIA-certified testing off-site if necessary
	Establish or expand institutional CLIA-compliant genotyping to expand point-of-care testing, same-day service
	Choose platform to assay multiple important genotypes simultaneously, reliably, and cheaply
	Invest in new equipment and personnel to ensure research quality control is at the same level as the clinical laboratory; this requires an institutional investment
	Work with genetic counselors to establish protocol for process and parameters of data return
	Develop better justification and approaches for reimbursing efforts of clinicians and genetic counselors
	Assess and provide sufficient genetic counselor staff or ancillary staff
	Establish institutional genomic medicine interdisciplinary clinical case conference for discussion and guidance on difficult clinical cases
Lack of standards for genomic applications	Develop agreed-upon framework or standards for evaluation of genomic medicine applications
	Develop standardized order sets and process modification
	Develop standards for analytic validity of whole-genome and whole-exome sequencing sufficient for clinical application, and for interpretation of the variants found by these methods
EMR integration of genomic results and clinical decision support	Enable access to actionable genetic information in the EMR through development of user-friendly decision support algorithms for health-care providers
	Establish "Usability Lab" to assess CDS tools and minimize alert fatigue
	Allocate genomic medicine institutional funding to develop education and outreach to disseminate "best practices" incorporating family history and genomic information
	Redesign the EMR to include section dedicated to containing all relevant genotype results for a patient
	Develop and link actionable drug–gene pair decision support to electronic pharmaceutical ordering software at point of care to facilitate adoption of pharmacogenomic testing
	Establish interdisciplinary workgroup with genomic medicine and EMR team to create secure tools for EMR-based genomic decision support
Follow-up of genotyped patients	Shift from relying on primary physician contact to direct contact and follow-up by genomic medicine team, with permission of physician
	Analyze and address reasons for refusal to complete confirmatory testing, such as lack of coverage for testing
Outreach to at-risk family members	Clarify implications for family members and health-care providers' responsibilities towards family members
	Explore ways to improve information to at-risk family members
Consent	Ensure consents for implementation projects conducted as research studies include returning results to patients and entering results into medical record
	Conformance with standard of care and specific consent may not be needed
	Consider initial implementation projects that might not require consent, such as results in established clinical pathways (tumor mutations, *CYP2C19*2*)
	Develop standards for informed consent for extensive sequencing, including whole-genome sequencing, and obtain it prospectively
	Ensure that availability of personnel to manage consent/counseling is not rate limiting in initial implementation

CDS, clinical decision support; CLIA, Clinical Laboratory Improvement Amendments; EMR, electronic medical record.

Table 2 Continued on next page

Table 2 Continued

Challenge	Potential solutions
Understanding by patients, clinicians, public	Conduct focus groups of patients, clinicians, and ancillary personnel to identify specific educational needs
	Conduct genetic and genomic medicine education campaign for patients, clinicians, and ancillary personnel based on focus group input
	Survey retention of educational information by patients and clinicians and modify programs as needed
	Conduct genotyping and/or sequencing and interpretation exercises with medical and other health profession students
	Provide specific health-care provider education on when to order test, and how to interpret and how to act on implications for family members and provider's responsibilities
	Introduce pharmacogenomic lectures into health professionals' training and continuing education
	Provide clinical supervision to clinician trainees in use of pharmacogenomic testing, other genomic point-of-care testing
Lack of access to comparison "control" sequence data and banking resources	Combine current centers' small patient collections of "normal" sequences and make available to all
	Prioritize funding for costly and time-consuming storage of viable tissues
	Bank high-quality tumor samples for confirmatory clinical sequencing with patient identifiers
	High-level institutional commitment to combine and organize multiple biorepositories for efficiency and ease of access
Lack of research funding and reimbursement	Until evidence base is established for making genomic testing new standard of care, consider research funding for testing in interim between discovery and adoption
	Gradually change culture to convince health-care community and patients of value of genomic medicine and need for reimbursement
	Demonstrate cost of testing is not prohibitive and savings and impact can be substantial
	Provide institutional back-up for reimbursement to avoid charges to patients
	Anticipate rises in interpretive and delivery costs as technology cost drops and enthusiasm increases

CDS, clinical decision support; CLIA, Clinical Laboratory Improvement Amendments; EMR, electronic medical record.

the Clinical Pharmacogenomics Implementation Consortium,[39] and by some early adopter institutions in developing committees of their own. Given the rapid spread of diagnostic sequencing and direct-to-consumer testing, it may be appropriate to begin to identify an intermediate category of "actionable" variants, for which evidence is insufficient to establish unequivocal clinical utility but is sufficient to determine how already available information could be used. Consideration should be given to tailoring of required levels of evidence to the risk/benefit ratio of the specific clinical setting at issue; development of consensus on evidence needed in different scenarios would be a valuable contribution. Recommendations for drastic or irreversible interventions such as genotype-driven mastectomy or salpingo–oophorectomy, for instance, should require far stricter evidence than should low-risk interventions such as modifying choice between two proven, roughly equivalent medications such as codeine and hydromorphone.[40]

For more modest interventions, the risks are more likely to involve factors such as burden to clinicians and patients in obtaining, interpreting, and managing the result, including learning of incidental findings not sought by clinician or patient; cost of testing; and psychological impact of the information on patients or their families. The first two of these can be addressed in part by improving the accessibility and lowering the costs of obtaining genomic information suitable for clinical use, which is occurring rapidly with expanding certification of genomic technologies and laboratories under CLIA and CAP. Testing fees can be substantial, especially when individual genetic variants are assayed sequentially, and indeed are soon likely to rival the cost of obtaining an entire genome sequence.[41] For a genetically heterogeneous condition, such as Charcot-Marie-Tooth disease, that is potentially due to variation at one or more of over 40 genetic loci, for example, whole-exome sequencing can already be performed for less than half the price of the current clinically available "multigene panel."[42] Third-party reimbursement of these costs is thus crucial for ensuring uptake by institutions and patients.

Integrating effective clinical decision support tools into the EMR, to query a patient's genomic data as care is being delivered, will be needed to present relevant variants and recommendations only when they are indicated. Such real-time, genomically driven clinical decision support can help avoid "alert fatigue" and relieve the clinician of trying to sift through the massive amount of genomic information without impeding clinical workflow. Sharing the effort of following up actionable genetic variants and at-risk family members with a dedicated genomic medicine team, with the primary physician's agreement and in ways that do not "compete" for the patient, can also minimize burden on patients and clinicians and increase efficiency. Looking to the future, the potential to update EMRs automatically as new discoveries are made and their clinical utility validated provides exciting possibilities for rapid implementation of research advances. Growing knowledge and understanding, however, mean that variants that are today of uncertain importance may tomorrow be found to be of great clinical relevance and vice versa. The responsibility of the clinician to recontact the patient with this new information is uncertain and could pose a major burden as well as medico-legal risks.[43]

Psychological impact on patients and their families has traditionally been addressed by the informed consent process and genetic counseling, but expanding genomic medicine efforts will likely soon outstrip the available supply of genetic counselors.[44] Innovative models utilizing ancillary personnel or social media tools to provide counseling and education, along with broad-ranging educational campaigns to increase familiarity with genomic medicine among practitioners, patients, and payers, will have the greatest impact if they can be linked to early successes and readily observable local efforts.

Reports of successes such as reduced hospitalization rates following genotype-driven adjustments in warfarin dosing,[45] e.g., or symptomatic improvement or reduced adverse effects after genotype-driven selection of antidepressants,[46] can help stimulate discussion of the potential for genomic medicine to improve outcomes and shift an institution's culture to embrace genotype-guided care,[30] although even these examples need a more complete evidence base and remain controversial.[47–50]

INFRASTRUCTURE AND RESEARCH NEEDS

Genomic medicine programs share needs for basic informational and policy infrastructure as well as evidence and outcomes produced through research (**Table 3**). Key infrastructure needs include a comprehensive knowledge base that captures sequence variants and their phenotypic associations and is openly accessible to clinical groups attempting to interpret sequence data. Such a knowledge base would have enormous value for research as well, particularly to the degree that it entails a careful curation and consensus process to reduce error and promote consistency. Here again, many genomic medicine groups appear to be doing this in isolation, capturing and reviewing such information internally but finding themselves constrained in sharing it outside their institutions. The policy aspects of establishing such a knowledge base, ensuring the accuracy of associations to be reported within it, and providing wide access to it are far from trivial and involve patient privacy and confidentiality of health information. Linking to appropriate supporting evidence and providing frequent updates are also daunting tasks, given the rapid pace of discovery, but are tasks that could readily be distributed among collaborating groups.

Perhaps an even greater informational need addressable by data sharing among genomic medicine groups and other genome sequencing projects is the ability to determine whether a newly discovered variant has been reported before. This is

Table 3 Infrastructure and research needs

Infrastructure
Continuously updated knowledge base and look-up tables of sequence variants linked to clinical phenotypes, genotype–phenotype relationships, actionable variants, and supportive evidence
Reliable, standardized genotyping/sequencing platforms, reporting formats, and quality control
High-throughput, rapid, low-cost genotyping/sequencing in CAP/CLIA-certified environments, potentially concentrated in large service centers for sequencing and interpretation
Evidentiary standards for various types of genomic testing in collaboration with other agencies (FDA, CMS) as needed, including standards and criteria for reimbursement
Point-of-care educational information and clinical decision support
Genomic medicine training programs for physicians, pharmacists, nurses, and other clinicians
Slide sets, videos, seminars, and social media approaches to patient, public, and clinician education
Increased person-power: clinicians with genetics training and background, genomic medicine consultants, genetic counselors, and implementation scientists
Expanded EMR capabilities to incorporate family history information, interface with genomic data
Research
Evidence development, such as changes following genomic medicine implementation in:
Disease severity, progression, or incidence
At-risk persons (such as family members) identified
Patients' or family members' anxiety, quality of life, and access to care or insurance
Cost, compliance, and patient or clinician satisfaction
Comparative effectiveness research
Identification of actionable variants
Bioinformatic/analysis tools to facilitate making associations, identifying "actionability" or impact scores for specific variants
Procedures and resources for interpreting potential clinical relevance of "variants of unknown significance," such as clinical proteomic tests, RNA-splicing tests, and cell-based assays, to determine experimentally whether a novel variant might cause disease
Standardized consent permitting return of results, deposition in EMR, sharing of variant tables
Collaborative projects/network to pool resources, share approaches, identify best practices, and avoid duplication
Best practices for:
Implementation and dissemination
Defining and collecting outcomes of implementation, including process outcomes
Delivering WGS information to individual providers and patients
Clearinghouse of successful implementation projects

CAP, College of American Pathologists; CLIA, Clinical Laboratory Improvement Amendments; CMS, Centers for Medicare and Medicaid Services; EMR, electronic medical record; FDA, Food and Drug Administration; WGS, whole-genome sequencing.

currently done by surveying available large-scale databases, such as the 1000 Genomes Project,[51] the Exome Variant Server,[52] and local institutional catalogs, on the presumption that a patient with a very rare or undiagnosed condition likely has a private or even unique variant.[53] Interpretation can be difficult because these reference databases generally lack phenotypic information. Far better than individual small cohorts of sequenced patients would be to aggregate cohorts and make their data widely available to qualified clinicians and researchers. Sequence variants in these databases should also be linked to phenotypic information, when possible, rather than forcing the assumption that all available sequenced individuals are free of the disease under consideration. The recently initiated Centers for Mendelian Genomics[54,55] will identify variants responsible for a great number of rare disorders and make this information widely available, which should help in this regard. The nascent ClinVar database of the National Center for Biotechnology Information[56] and the International Standards for Cytogenomic Arrays database available in dbGaP and dbVar[57] are two open-access databases that can provide valuable phenotypic information associated with rare sequence variants and structural variation.

Placing all this information in a single queryable source, regardless of where the actual data reside, would be a substantial advance in interpreting the three million or more differences from the human reference sequence identifiable per individual.[58] Genomic sequencing, in contrast to genotyping microarray technology, reveals rare or even unique alleles lacking scientific literature support for physiological effects. Computational approaches to predicting function may need to be incorporated into clinical decision support methods in trying to determine the implications of such variants for clinical care. Such analyses would be greatly facilitated by standard formats for reporting the data produced by individual sequencing groups rather than expecting individual clinicians and clinical sites to handle and interpret multiple different formats. The Centers for Disease Control and Prevention are currently working to develop performance specifications for next-generation sequencing.[59] Access to accurate, low-cost genome sequencing in CAP/CLIA-certified environments will remain a major limiting factor in utilizing sequencing in clinical care across diverse settings. Provision of genome sequencing as a commodity service in national or regional hubs, including interpretation of potential clinical relevance of identified variants, could be a major advance.

Appropriate professional organizations that can develop genomic medicine practice guidelines, such as the American College of Medical Genetics and Genomics, the College of American Pathologists, and the Association of Molecular Pathologists, as well as regulatory agencies such as the FDA and Centers for Medicare and Medicaid Services, should be involved in developing evidentiary standards tailored to specific interventions under consideration and criteria for reimbursement. Targeted education for active practitioners should also be tailored to specific settings and delivered as succinctly as possible at the point of care for maximum uptake and value, whereas broader educational efforts can capture clinicians earlier in their training.[60] Fellowships in pharmacogenomics and genomic medicine will be needed to develop specialists (perhaps "clinical genomicists," akin to radiologists or pathologists) capable of interpreting patients' genomic information and advising clinicians on appropriate actions to be taken for a given set of variants in a given clinical setting. Given the broad range of possible genomic medicine applications, the training of physicians and scientists enrolled in certified programs of the American Board of Medical Genetics and Genomics and the American Board of Pathology in clinical genetics, molecular genetics, biochemical genetics, and cytogenetics is particularly relevant. Consideration should be given to the potential for expanding the curricula of these programs to include more genomic medicine content or providing fellowships within these specialties that focus on genomic medicine.

Other educational modes such as seminars, slide sets, webinars, and videos need to be developed and shared across institutions doing similar types of work. Expanded EMR capabilities to permit querying of sequence or other dense genomic data, and incorporation of family history, are readily addressable. EMRs also have the ability to deliver context-specific educational information at the point of care during an eminently teachable moment, with a clinical conundrum at hand, as well as more sophisticated clinical decision support to aid clinicians in the appropriate use of genomic information.[61] Challenges in implementing point-of-care tools need to be recognized, however, including the potential for overwhelming clinicians with alerts leading to alert fatigue.[62] Here again, consideration should be given to the evidence needed for implementing such tools in genomic medicine and for ensuring the consistent adoption of point-of-care interventions for high priority health-care targets with the best risk/benefit ratio and strongest evidence, whether genomic or nongenomic.

Research is acutely needed to generate, collect, and make widely available the evidence needed for determining which variants are actionable, in whom, and in what clinical situations. Intermediate steps are needed that lie between genomic discovery research and routine clinical implementation to generate the evidence base to justify implementation. Consensus should be sought on the levels of evidence needed to justify the introduction of genomic medicine interventions, taking into account the degree of complexity or cost (such as family history vs. whole-genome sequencing) and risk/benefit (Lynch syndrome screening vs. universal colonscopy) involved, and the comparisons and outcome measures needed to answer these questions. Generating clear evidence of benefit for genomic interventions will remain challenging, however, in part because the rarity of many genomic markers of risk means large samples must be studied to accrue benefit/risk data in the small subsets of patients affected. Even with large samples, there can be controversy about the interpretation of the evidence, as shown by the data on the relationship between *CYP2C19* variants and failure of antiplatelet efficacy of clopidogrel.[63] In one meta-analysis that included only patients with acute coronary syndromes, there was strong evidence for an effect of variant alleles on outcomes,[23] whereas another analysis that included trials with different indications (and in which the efficacy of clopidogrel itself appears to be less) reported a much smaller effect of genetic variation on outcome.[24] Thus, the evidence supporting an effect of pharmacogenomic variants on variable drug outcomes may also depend on indication, substantially increasing the complexity of generating and evaluating evidence.

Research efforts should also aim to establish and validate criteria for actionability. They should facilitate association analysis and assessment of actionability and influence on outcomes from available data. Given that it is impossible, however, to perform prospective studies of the impact of all potentially important variants on clinical outcomes, comprehensive strategies will be needed to generate the necessary data. Real-world studies quantifying the impact of implementing genomic medicine programs on patient outcomes, cost and acceptability of care, and clinician and patient satisfaction are needed to promote adoption and dissemination of genomic medicine. Such studies fall under the rubric of "implementation research," defined as the study of methods to promote the systematic uptake of proven interventions into routine practice.[64] Comparative effectiveness research assessing the added value of genomic information, such as genomic variants added to family history added to clinical risk factors, and consensus on the outcomes and evidence needed to answer such questions, would help to define priorities for collecting and using such information in a cost-effective way.

Informed consents for genomic research should be broadened to include return of results, where desired and appropriate, so that lack of such consent (or indeed, consent forms declaring "no results will ever be returned") does not become an obstacle, if not an excuse, for failing to report actionable findings. The consent process should also explicitly address deposition of data in the EMR or research databases and sharing of variant information among researchers and clinicians.

Assembling early adopters of genomic medicine into a confederation or community should facilitate addressing the many infrastructure and research needs outlined here. Collaborative projects such as those of the PGRN[65] can facilitate dissemination and testing of genomic medicine approaches in diverse settings to maximize their generalizability and usefulness. Such projects can help develop best practices for complex tasks such as implementation and dissemination, defining and collecting outcomes of implementation, and delivering genomic information to clinicians and patients. A clearinghouse of successful implementation projects, with detailed protocols addressing steps needed for patients, clinicians, laboratories, departments, and institutions, would go a long way toward disseminating this work beyond early adopter sites. Indeed, a critical question to ask is whether and how approaches developed at highly specialized and resourced tertiary care centers can be adopted in less resource-intensive settings.

IMPLEMENTATION ROADMAP

Incorporating genomic results into clinical care is as much a cultural and political exercise within a given institution as a scientific one, although a firm scientific grounding is essential to convincing the various stakeholders to accept an implementation project, particularly the first one (**Figure 1**). Such grounding can be obtained from a comprehensive review of the available scientific literature, recommendations of expert groups, and examination of ongoing successful projects at other institutions. Available policy or institutional guidelines on implementing changes to clinical care must also be considered. Critical in choosing an initial implementation project is the pragmatic issue of identifying a group of ready adopters or clinical champions, such as a subspecialty group or local practice, willing to try a new approach. Indeed, selection of a pilot project among several possibilities may hinge upon the availability of an enthusiastic group of clinicians.

Identifying and engaging the myriad stakeholders within an institution, including those needed to conduct genetic testing, interpret it, integrate it with the EMR, provide results and recommendations to the clinician, and pay for it, require that senior leaders who are familiar with the institution and its culture be on the genomic medicine team. It also often requires patience, persistence, and sufficient standing within the organization to convince skeptics and overcome institutional inertia. Some programs have found it simpler to begin at smaller affiliated hospitals or practices rather than at dominant academic medical centers.

Engagement and support from institutional leadership is obviously crucial, and most successful programs have involved

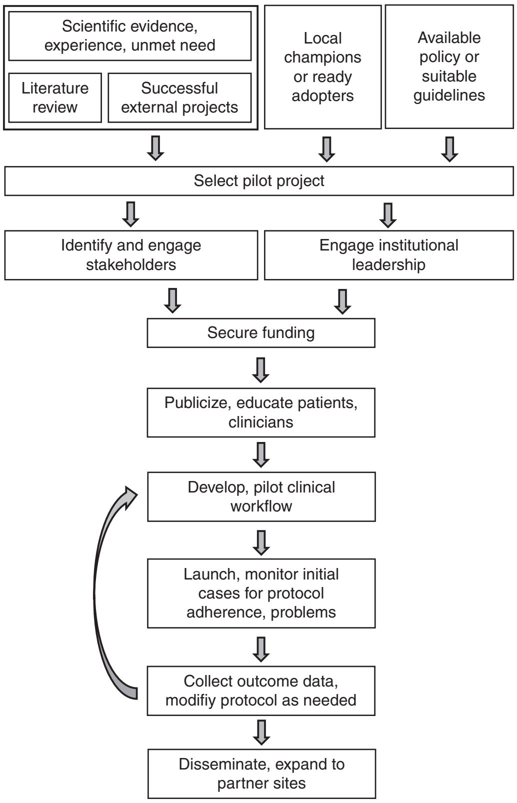

Scientific evidence, experience, unmet need
Literature review
Successful external projects
Local champions or ready adopters
Available policy or suitable guidelines
Select pilot project
Identify and engage stakeholders
Engage institutional leadership
Secure funding
Publicize, educate patients, clinicians
Develop, pilot clinical workflow
Launch, monitor initial cases for protocol adherence, problems
Collect outcome data, modifiy protocol as needed
Disseminate, expand to partner sites

Figure 1 Implementation roadmap: one approach to initial implementation at a single institution with subsequent evaluation and ultimately dissemination.

them directly in decision making, as chairs of committees selecting interventions to implement, for example, or evaluating their impact. Funding may be provided through local institutional support, in anticipation perhaps of eventual reimbursement by major payers, or through foundation or research funding, including National Institutes of Health–supported projects such as the Clinical Translational Science Awards and the PGRN. Institutional stakeholders are key in developing practical and efficient clinical workflows for obtaining testing and providing results. Development and testing of streamlined educational tools, which may be as simple as a pop-up window with appropriate links for the clinician and letters or brochures for patients, will be essential both for initiating a program and encouraging its acceptance. Once launched, careful monitoring of initial cases for adherence to established workflows and unanticipated barriers permits tailoring of procedures as needed. Outcome data related to ease of use, adherence, patient and clinician satisfaction and behaviors, cost, and, where possible, measures of morbidity and mortality can be used to modify the program and inform design and implementation of future efforts. Evaluation of the effectiveness of an implementation project should focus on outcomes of the implementation itself rather than the outcomes of treatment.[66] Several conceptual models for such evaluations are available for consideration.[67,68]

CONCLUSION

Multiple institutions are actively engaged in using patients' genomic information in their clinical care, either as pilot demonstration and research projects or as part of clinical care as a reimbursable service. The Genomic Medicine Colloquium revealed that much of this work is being done in isolation and would benefit from more structured collaboration and sharing of approaches across groups and institutions. Implementation of genomics in medicine would also benefit from establishing common infrastructure, such as a catalog of variants observed across large numbers of sequenced individuals to date, an updatable database of actionable variants, evidentiary standards tailored to benefits and risks, common clinical decision support tools including point-of-care educational materials, and collaborative projects to pool resources and identify best practices. Systematic and timely incorporation of advances in genomic research related to somatic and germline DNA, as well as consideration of RNA, protein, and epigenomic applications, will also be needed. The National Human Genome Research Institute, in collaboration with several other National Institutes of Health institutes and in accord with its recent strategic plan,[69] is currently exploring approaches for facilitating the research needed to implement genomic medicine on an ever-widening scale. Such efforts should also take full advantage of the extant momentum, information, and experience of "early adopters" across the United States, as well as the tools of implementation science to address the cultural barriers described above.

The substantial National Institutes of Health investment in basic genomic research has yielded and will continue to yield new insights into the role of genetic variants in health and disease. It is now time to ensure those results are translated and used to the maximum benefit of patients by periodically convening relevant stakeholders and building on the vanguard of nationwide genomic medicine implementation projects. Bringing these groups together in this way is anticipated to increase their diffusion, efficiency, and, ultimately, their value to patients and health care.

ACKNOWLEDGMENTS

This article summarizes the deliberations of a colloquium convened by the National Human Genome Research Institute on 29 June 2011 to examine opportunities and barriers for implementing genomic medicine in clinical care. The authors gratefully acknowledge the contributions of Gary Gibbons to the workshop.

DISCLOSURE

C.E. is an unpaid member of the External Scientific Advisory Board of MyMercury.com and of the Genomic Medicine Advisory Board of Complete Genomics, Inc. The other authors declare no conflict of interest.

REFERENCES

1. Dulbecco R. A turning point in cancer research: sequencing the human genome. *Science* 1986;231:1055–1056.
2. Collins FS. Shattuck lecture–medical and societal consequences of the Human Genome Project. *N Engl J Med* 1999;341:28–37.
3. Guttmacher AE, Collins FS. Genomic medicine–a primer. *N Engl J Med* 2002;347:1512–1520.
4. Evans JP, Meslin EM, Marteau TM, Caulfield T. Genomics. Deflating the genomic bubble. *Science* 2011;331:861–862.
5. Varmus H. Ten years on–the human genome and medicine. *N Engl J Med* 2010;362:2028–2029.
6. Committee on Quality of Health Care in America, Institute of Medicine. *Crossing the Quality Chasm: A New Health System for the 21st Century*. National Academy Press: Washington, DC, 2001.
7. England JD, Gronseth GS, Franklin G, et al.; American Academy of Neurology. Practice Parameter: evaluation of distal symmetric polyneuropathy: role of laboratory and genetic testing (an evidence-based review). Report of the American Academy of Neurology, American Association of Neuromuscular and Electrodiagnostic Medicine, and American Academy of Physical Medicine and Rehabilitation. *Neurology* 2009;72:185–192.
8. Lupski JR, de Oca-Luna RM, Slaugenhaupt S, et al. DNA duplication associated with Charcot-Marie-Tooth disease type 1A. *Cell* 1991;66:219–232.
9. Manolio TA, Brooks LD, Collins FS. A HapMap harvest of insights into the genetics of common disease. *J Clin Invest* 2008;118:1590–1605.
10. Ladabaum U, Wang G, Terdiman J, et al. Strategies to identify the Lynch syndrome among patients with colorectal cancer: a cost-effectiveness analysis. *Ann Intern Med* 2011;155:69–79.
11. Coates R, Williams M, Melillo S, Gudgeon J. Genetic testing for lynch syndrome in individuals newly diagnosed with colorectal cancer to reduce morbidity and mortality from colorectal cancer in their relatives. *PLoS Curr* 2011;3:RRN1246.
12. US Preventive Services Task Force. Genetic risk assessment and BRCA mutation testing for breast and ovarian cancer susceptibility: recommendation statement. *Ann Intern Med* 2005;143:355–361.
13. Gudgeon JM, Williams JL, Burt RW, Samowitz WS, Snow GL, Williams MS. Lynch syndrome screening implementation: business analysis by a healthcare system. *Am J Manag Care* 2011;17:e288–e300.
14. Qureshi N, Armstrong S, Dhiman P, et al.; for the ADDFAM (Added Value of Family History in CVD Risk Assessment) Study Group. Effect of adding systematic family history enquiry to cardiovascular disease risk assessment in primary care: A Matched-Pair, Cluster Randomized Trial. *Ann Intern Med* 2012;156:253–262.
15. Rubinstein WS, Acheson LS, O'Neill SM, et al.; Family Healthware Impact Trial (FHITr) Group. Clinical utility of family history for cancer screening and referral in primary care: a report from the Family Healthware Impact Trial. *Genet Med* 2011;13:956–965.

16. Khoury MJ, Feero WG, Valdez R. Family history and personal genomics as tools for improving health in an era of evidence-based medicine. *Am J Prev Med* 2010;39:184–188.
17. Surgeon General's My Family Health Portrait. https://familyhistory.hhs.gov/fhh-web/home.action.
18. Orlando LA, Hauser ER, Christianson C, et al. Protocol for implementation of family health history collection and decision support into primary care using a computerized family health history system. *BMC Health Serv Res* 2011;11:264.
19. Doerr M, Leach B, Men X, et al. Patient-entered, web-based, EMR-compatible, comprehensive cancer family history collection tool: clinical workflow, demographics, and patient satisfaction from pilot study. American Society of Human Genetics 2010 Annual Meeting: Abstract 1992. Washington, D.C., 2–6 November 2010.
20. Hulse NC, Ranade-Kharkar P, Post H, Wood GM, Williams MS, Haug PJ. Development and early usage patterns of a consumer-facing family health history tool. *AMIA Annu Symp Proc* 2011;2011:578–587.
21. Pulley JM, Denny JC, Peterson JF, et al. Operational implementation of prospective genotyping for personalized medicine: the design of the Vanderbilt PREDICT project. *Clin Pharmacol Ther* 2012;92:87–95.
22. Collet JP, Hulot JS, Pena A, et al. Cytochrome P450 2C19 polymorphism in young patients treated with clopidogrel after myocardial infarction: a cohort study. *Lancet* 2009;373:309–317.
23. Mega JL, Simon T, Collet JP, et al. Reduced-function CYP2C19 genotype and risk of adverse clinical outcomes among patients treated with clopidogrel predominantly for PCI: a meta-analysis. *JAMA* 2010;304:1821–1830.
24. Holmes MV, Perel P, Shah T, Hingorani AD, Casas JP. CYP2C19 genotype, clopidogrel metabolism, platelet function, and cardiovascular events: a systematic review and meta-analysis. *JAMA* 2011;306:2704–2714.
25. St. Jude Children's Research Hospital PG4KDS program. http://www.stjude.org/pg4kds.
26. Pharmacogenomics Research Network (PGRN). http://www.nigms.nih.gov/Research/FeaturedPrograms/PGRN/.
27. Pharmacogenomics Knowledge Base (PharmGKB). http://www.pharmgkb.org/search/guidelineList.action.
28. US Food and Drug Administration. Table of Pharmacogenomic Biomarkers in Drug Labels. http://www.fda.gov/drugs/scienceresearch/researchareas/pharmacogenetics/ucm083378.htm.
29. McCarty CA, Chisholm RL, Chute CG, et al.; eMERGE Team. The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Med Genomics* 2011;4:13.
30. Gonzaga-Jauregui C, Lupski JR, Gibbs RA. Human genome sequencing in health and disease. *Annu Rev Med* 2012;63:35–61.
31. Worthey EA, Mayer AN, Syverson GD, et al. Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. *Genet Med* 2011;13:255–262.
32. St Hilaire C, Ziegler SG, Markello TC, et al. NT5E mutations and arterial calcifications. *N Engl J Med* 2011;364:432–442.
33. Baylor College of Medicine, Whole Genome Laboratory. http://www.bcm.edu/geneticlabs/index.cfm?PMID=21319.
34. Ambry Genetics. http://www.ambrygen.com/clinical-diagnostic-exome.
35. GeneDx. http://www.genedx.com/site/xomedx.
36. SeqWright. http://www.seqwright.com/Exome.
37. Scheuner MT, Sieverding P, Shekelle PG. Delivery of genomic medicine for common chronic adult diseases: a systematic review. *JAMA* 2008;299:1320–1334.
38. Teutsch SM, Bradley LA, Palomaki GE, et al.; EGAPP Working Group. The Evaluation of Genomic Applications in Practice and Prevention (EGAPP) Initiative: methods of the EGAPP Working Group. *Genet Med* 2009;11:3–14.
39. Relling MV, Klein TE. CPIC: Clinical Pharmacogenetics Implementation Consortium of the Pharmacogenomics Research Network. *Clin Pharmacol Ther* 2011;89:464–467.
40. Crews KR, Gaedigk A, Dunnenberger HM, et al.; Clinical Pharmacogenetics Implementation Consortium. Clinical Pharmacogenetics Implementation Consortium (CPIC) guidelines for codeine therapy in the context of cytochrome P450 2D6 (CYP2D6) genotype. *Clin Pharmacol Ther* 2012;91:321–326.
41. Salzberg SL, Pertea M. Do-it-yourself genetic testing. *Genome Biol* 2010;11:404.
42. Lupski JR, Reid JG, Gonzaga-Jauregui C, et al. Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. *N Engl J Med* 2010;362:1181–1191.
43. Pyeritz RE. The coming explosion in genetic testing–is there a duty to recontact? *N Engl J Med* 2011;365:1367–1369.
44. National Cancer Institute, Research-tested Intervention Programs. http://rtips.cancer.gov/rtips/programDetails.do?programId=277363.
45. Wang L, McLeod HL, Weinshilboum RM. Genomics and drug response. *N Engl J Med* 2011;364:1144–1153.
46. Porcelli S, Drago A, Fabbri C, Gibiino S, Calati R, Serretti A. Pharmacogenetics of antidepressant response. *J Psychiatry Neurosci* 2011;36:87–113.
47. Verhoef TI, Schalekamp T, Redekop WK, de Boer A, Maitland-van der Zee AH. Clinical and economic consequences of pharmacogenetic-guided dosing of warfarin. *Expert Rev Pharmacoecon Outcomes Res* 2010;10:375–378.
48. Meckley LM, Gudgeon JM, Anderson JL, Williams MS, Veenstra DL. A policy model to evaluate the benefits, risks and costs of warfarin pharmacogenomic testing. *Pharmacoeconomics* 2010;28:61–74.
49. Evaluation of Genomic Applications in Practice and Prevention (EGAPP) Working Group. Recommendations from the EGAPP Working Group: testing for cytochrome P450 polymorphisms in adults with nonpsychotic depression treated with selective serotonin reuptake inhibitors. *Genet Med* 2007;9:819–825.
50. Anderson JL, Horne BD, Stevens SM, et al. A randomized and clinical effectiveness trial comparing two pharmacogenetic algorithms and standard care for individualizing warfarin dosing (CoumaGen-II). *Circulation* 2012;125:1997–2005.
51. 1000 Genomes Project. http://www.1000genomes.org/.
52. Exome Variant Server. http://evs.gs.washington.edu/EVS/.
53. Lupski JR, Belmont JW, Boerwinkle E, Gibbs RA. Clan genomics and the complex architecture of human disease. *Cell* 2011;147:32–43.
54. Mendelian Disorders Sequencing Centers. http://www.genome.gov/27546192.
55. Bamshad MJ, Shendure JA, Valle D, et al.; Centers for Mendelian Genomics. The Centers for Mendelian Genomics: a new large-scale initiative to identify the genes underlying rare Mendelian conditions. *Am J Med Genet A* 2012;158A:1523–1525.
56. National Center for Biotechnology Information. http://www.ncbi.nlm.nih.gov/clinvar/.
57. International Standards for Cytogenomic Arrays (ISCA) database. http://www.iscaconsortium.org.
58. Bainbridge MN, Wiszniewski W, Murdock DR, et al. Whole-genome sequencing for optimized patient management. *Sci Transl Med* 2011;3:87re3.
59. Centers for Disease Control and Prevention, Next Generation Sequencing: Standardization of Clinical Testing (Nex-StoCT) Working Group. http://www.cdc.gov/osels/lspppo/Genetic_Testing_Quality_Practices/Nex-StoCT.html.
60. Dhar SU, Alford RL, Nelson EA, Potocki L. Enhancing exposure to genetics and genomics through an innovative medical school curriculum. *Genet Med* 2012;14:163–167.
61. Hoffman MA, Williams MS. Electronic medical records and personalized medicine. *Hum Genet* 2011;130:33–39.
62. Shah NR, Seger AC, Seger DL, et al. Improving acceptance of computerized prescribing alerts in ambulatory care. *J Am Med Inform Assoc* 2006;13:5–11.
63. Johnson JA, Roden DM, Lesko LJ, Ashley E, Klein TE, Shuldiner AR. Clopidogrel: a case for indication-specific pharmacogenetics. *Clin Pharmacol Ther* 2012;91:774–776.
64. Eccles MP, Foy R, Sales A, Wensing M, Mittman B. Implementation Science six years on–our evolving scope and common reasons for rejection without review. *Implement Sci* 2012;7:71.
65. Pharmacogenomics Research Network Research Groups. http://pgrn.org/display/pgrnwebsite/Research+Groups+and+Network+Resources.
66. Proctor E, Silmere H, Raghavan R, et al. Outcomes for implementation research: conceptual distinctions, measurement challenges, and research agenda. *Adm Policy Ment Health* 2011;38:65–76.
67. Proctor EK, Landsverk J, Aarons G, Chambers D, Glisson C, Mittman B. Implementation research in mental health services: an emerging science with conceptual, methodological, and training challenges. *Adm Policy Ment Health* 2009;36:24–34.
68. Gaglio B, Glasgow RE. Evaluation approaches for dissemination and implementation research. In: Brownson RC, Colditz GA, Proctor EK (eds.), *Dissemination and Implementation Research in Health: Translating Science to Practice*. Oxford University Press, New York, NY 2012, pp. 334–337.
69. Green ED, Guyer MS; National Human Genome Research Institute. Charting a course for genomic medicine from base pairs to bedside. *Nature* 2011;470:204–213.

This work is licensed under a Creative Commons Attribution-NonCommercial-No Derivative Works 3.0 License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-nd/3.0/

medicina

MDPI

Commentary

Universal Healthcare in the United States of America: A Healthy Debate

Gabriel Zieff [1,*], Zachary Y. Kerr [1], Justin B. Moore [2] and Lee Stoner [1]

[1] Department of Exercise and Sport Science, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA; zkerr@email.unc.edu (Z.Y.K.); dr.l.stoner@gmail.com (L.S.)
[2] Department of Implementation Science, Wake Forest School of Medicine, Winston-Salem, NC 27157, USA; jusmoore@wakehealth.edu
* Correspondence: gzieff@live.unc.edu; Tel.: +1-(919)-962-5755

Received: 6 September 2020; Accepted: 28 October 2020; Published: 30 October 2020

Abstract: This commentary offers discussion on the pros and cons of universal healthcare in the United States. Disadvantages of universal healthcare include significant upfront costs and logistical challenges. On the other hand, universal healthcare may lead to a healthier populace, and thus, in the long-term, help to mitigate the economic costs of an unhealthy nation. In particular, substantial health disparities exist in the United States, with low socio–economic status segments of the population subject to decreased access to quality healthcare and increased risk of non-communicable chronic conditions such as obesity and type II diabetes, among other determinants of poor health. While the implementation of universal healthcare would be complicated and challenging, we argue that shifting from a market-based system to a universal healthcare system is necessary. Universal healthcare will better facilitate and encourage sustainable, preventive health practices and be more advantageous for the long-term public health and economy of the United States.

Keywords: chronic disease; health insurance; socio–economic status; obesity; diabetes; hypertension; health promotion; universal healthcare

1. Introduction

Healthcare is one of the most significant socio–political topics in the United States (U.S.), and citizens currently rank "healthcare" as the most important issue when it comes to voting [1]. The U.S. has historically utilized a mixed public/private approach to healthcare. In this approach, citizens or businesses can obtain health insurance from private (e.g., Blue Cross Blue Shield, Kaiser Permanente) insurance companies, while individuals may also qualify for public (e.g., Medicaid, Medicare, Veteran's Affairs), government-subsidized health insurance. In contrast, the vast majority of post-industrial, Westernized nations have used various approaches to provide entirely or largely governmentally subsidized, universal healthcare to all citizens regardless of socio–economic status (SES), employment status, or ability to pay. The World Health Organization defines universal healthcare as "ensuring that all people have access to needed health services (including prevention, promotion, treatment, rehabilitation and palliation) of sufficient quality to be effective while also ensuring that the use of these services does not expose the user the financial hardship" [2]. Importantly, the Obama-era passage of the Affordable Care Act (ACA) sought to move the U.S. closer to universal healthcare by expanding health coverage for millions of Americans (e.g., via Medicaid expansion, launch of health insurance marketplaces for private coverage) including for citizens across income levels, age, race, and ethnicity.

Differing versions of universal healthcare are possible. The United Kingdom's National Health Services can be considered a fairly traditional version of universal healthcare with few options for, and minimal use of, privatized care [3]. On the other hand, European countries like Switzerland, the Netherlands, and Germany have utilized a blended system with substantial government and market-based components [4,5]. For example, Germany uses a multi-payer healthcare system in which subsidized health care is widely available for low-income citizens, yet private options—which provide the same quality and level of care as the subsidized option—are also available to higher income individuals. Thus, universal healthcare does not necessarily preclude the role of private providers within the healthcare system, but rather ensures that equity and effectiveness of care at population and individual levels are a reference and expectation for the system as a whole. In line with this, versions of universal healthcare have been implemented by countries with diverse political backgrounds (e.g., not limited to traditionally "socialist/liberal" countries), including some with very high degrees of economic freedom [6,7].

Determining the degree to which a nation's healthcare is "universal" is complex and is not a "black and white" issue. For example, government backing, public will, and basic financing structure, among many other factors must be extensively considered. While an in-depth analysis of each of these factors is beyond the scope of this commentary, there are clear advantages and disadvantages to purely private, market-based, and governmental, universal approaches to healthcare, as well as for policies that lie somewhere in-between. This opinion piece will highlight arguments for and against universal healthcare in the U.S., followed by the authors' stance on this issue and concluding remarks.

2. Argument against Universal Healthcare

Though the majority of post-industrial Westernized nations employ a universal healthcare model, few—if any—of these nations are as geographically large, populous, or ethnically/racially diverse as the U.S. Different regions in the U.S. are defined by distinct cultural identities, citizens have unique religious and political values, and the populace spans the socio–economic spectrum. Moreover, heterogenous climates and population densities confer different health needs and challenges across the U.S. [8]. Thus, critics of universal healthcare in the U.S. argue that implementation would not be as feasible—organizationally or financially—as other developed nations [9]. There is indeed agreement that realization of universal healthcare in the U.S. would necessitate significant upfront costs [10]. These costs would include those related to: (i) physical and technological infrastructural changes to the healthcare system, including at the government level (i.e., federal, state, local) as well as the level of the provider (e.g., hospital, out-patient clinic, pharmacy, etc.); (ii) insuring/treating a significant, previously uninsured, and largely unhealthy segment of the population; and (iii) expansion of the range of services provided (e.g., dental, vision, hearing) [10].

The cost of a universal healthcare system would depend on its structure, benefit levels, and extent of coverage. However, most proposals would entail increased federal taxes, at least for higher earners [4,11,12]. One proposal for universal healthcare recently pushed included options such as a 7.5% payroll tax plus a 4% income tax on all Americans, with higher-income citizens subjected to higher taxes [13]. However, outside projections suggest that these tax proposals would not be sufficient to fund this plan. In terms of the national economic toll, cost estimations of this proposal range from USD 32 to 44 trillion across 10 years, while deficit estimations range from USD 1.1 to 2.1 trillion per year [14].

Beyond individual and federal costs, other common arguments against universal healthcare include the potential for general system inefficiency, including lengthy wait-times for patients and a hampering of medical entrepreneurship and innovation [3,12,15,16]. Such critiques are not new, as exemplified by rhetoric surrounding the Clinton Administration's Health Security Act which was labeled as "government meddling" in medical care that would result in "big government inefficiency" [12,15]. The ACA has been met with similar resistance and bombast (e.g., the "repeal and replace" right-leaning rallying cry) as a result of perceived inefficiency and unwanted government involvement. As an example of lengthy wait times associated with universal coverage, in 2017 Canadians were on waiting lists for an estimated 1,040,791 procedures, and the median wait time for arthroplastic surgery was 20–52 weeks [17]. Similarly, average waiting time for elective hospital-based care in the United Kingdom is 46 days, while some patients wait over a year (3). Increased wait times in the U.S. would likely occur—at least in the short term—as a result of a steep rise in the number of primary and emergency care visits (due to eliminating the financial barrier to seek care), as well as general wastefulness, inefficiency, and disorganization that is often associated with bureaucratic, government-run agencies.

3. Argument for Universal Healthcare

Universal healthcare in the U.S., which may or may not include private market-based options, offer several noteworthy advantages compared to exclusive systems with inequitable access to quality care including: (i) addressing the growing chronic disease crisis; (ii) mitigating the economic costs associated with said crisis; (iii) reducing the vast health disparities that exist between differing SES segments of the population; and (iv) increasing opportunities for preventive health initiatives [18–21]. Perhaps the most striking advantage of a universal healthcare system in the U.S. is the potential to address the epidemic level of non-communicable chronic diseases such as cardiovascular diseases, type II diabetes, and obesity, all of which strain the national economy [22,23]. The economic strain associated with an unhealthy population is particularly evident among low SES individuals. Having a low SES is associated with many unfavorable health determinants, including decreased access to, and quality of health insurance which impact health outcomes and life expectancies [24]. Thus, the low SES segments of the population are in most need of accessible, quality health insurance, and economic strain results from an unhealthy and uninsured low SES [25,26]. For example, diabetics with low SES have a greater mortality risk than diabetics with higher SES, and the uninsured diabetic population is responsible for 55% more emergency room visits each year than their insured diabetic counterparts [27,28]. Like diabetes, hypertension—the leading risk factor for death worldwide [29], has a much higher prevalence among low SES populations [30]. It is estimated that individuals with uncontrolled hypertension have more than USD 2000 greater annual healthcare costs than their normotensive counterparts [31]. Lastly, the incidence of obesity is also much greater among low SES populations [32]. The costs of obesity in the U.S., when limited to lost productivity alone, have been projected to equate to USD 66 billion annually [33]. Accessible, affordable healthcare may enable earlier intervention to prevent—or limit risk associated with—non-communicable chronic diseases, improve the overall public health of the U.S., and decrease the economic strain associated with an unhealthy low-SES.

Preventive Initiatives within A Universal Healthcare Model

Beyond providing insurance coverage for a substantial, uninsured, and largely unhealthy segment of society—and thereby reducing disparities and unequal access to care among all segments of the population—there is great potential for universal healthcare models to embrace value-based care [4,20,34]. Value-based care can be thought of as appropriate and affordable care (tackling wastes), and integration of services and systems of care (i.e., hospital, primary, public health), including preventive care that considers the long-term health and economy of a nation [34,35]. In line with this, the ACA has worked in parallel with population-level health programs such as the Healthy People Initiative by targeting modifiable determinants of health including physical activity, obesity, and environmental quality, among others [36]. Given that a universal healthcare plan would force the government to pay for costly care and treatments related to complications resulting from preventable, non-communicable chronic diseases, the government may be more incentivized to (i) offer primary prevention of chronic disease risk prior to the onset of irreversible complications, and (ii) promote wide-spread preventive efforts across multiple societal domains. It is also worth acknowledging here that the national public health response to the novel Coronavirus-19 virus is a salient and striking contemporary example of a situation in which there continues to be a need to expeditiously coordinate multiple levels of policy, care, and prevention.

Preventive measures lessen costs associated with an uninsured and/or unhealthy population [37]. For example, investing USD 10 per person annually in community-based programs aimed at combatting physical inactivity, poor nutrition, and smoking in the U.S. could save more than USD 16 billion annually within five years, equating to a return of USD 5.60 for every dollar spent [38]. Another recent analysis suggests that if 18% more U.S. elementary-school children participated in 25 min of physical activity three times per week, savings attributed to medical costs and productivity would amount to USD 21.9 billion over their lifetime [39]. Additionally, simple behavioral changes can have major clinical implications. For example, simply brisk walking for 30 min per day (≥15 MET-hours/week) has been associated with a 50% reduction in type II diabetes [40]. While universal healthcare does not necessarily mean that health policies supporting prevention will be enacted, it may be more likely to promote healthy (i) lifestyle behaviors (e.g., physical activity), (ii) environmental factors (e.g., safe, green spaces in low and middle-income communities), and (iii.) policies (e.g., banning sweetened beverages in public schools) compared to a non-inclusive system [34–36].

Nordic nations provide an example of inclusive healthcare coupled with multi-layered preventive efforts [41]. In this model, all citizens are given the same comprehensive healthcare while social determinants of health are targeted. This includes "mobilizing and coordinating a large number of players in society," which encourages cooperation among "players" including municipal political bodies, voluntary organizations, and educational institutions [41]. Developmental and infrastructural contributions from multiple segments of society to a healthcare system may also better encourage government accountability compared to a system in which a select group of private insurers and citizens are the only "stakeholders." Coordinated efforts on various non-insurance-related fronts have focused on obesity, mental health, and physical activity [41]. Such coordinated efforts within the Nordic model have translated to positive health outcomes. For example, the Healthcare Access and Quality (HAQ) Index provides an overall score of 0–100 (0 being the worst) for healthcare access and quality across 195 countries and reflects rates of 32 preventable causes of death. Nordic nations had an average HAQ score of 95.4, with four of the five nations achieving scores within the top 10 worldwide [42]. Though far more heterogenous compared to Nordic nations, (e.g., culturally, geographically, racially, etc.), the U.S. had a score of 89 (29th overall) [42]. To provide further context, other industrialized nations, which are more comparable to the U.S. than Nordic nations, also ranked higher than the U.S. including Germany (92, 19th overall), Canada (94, 14th overall), Switzerland (96, 7th overall), and the Netherlands (96, 3rd overall) [42].

4. Conclusions

Non-inclusive, inequitable systems limit quality healthcare access to those who can afford it or have employer-sponsored insurance. These policies exacerbate health disparities by failing to prioritize preventive measures at the environmental, policy, and individual level. Low SES segments of the population are particularly vulnerable within a healthcare system that does not prioritize affordable care for all or address important determinants of health. Failing to prioritize comprehensive, affordable health insurance for all members of society and straying further from prevention will harm the health and economy of the U.S. While there are undoubtedly great economic costs associated with universal healthcare in the U.S., we argue that in the long-run, these costs will be worthwhile, and will eventually be offset by a healthier populace whose health is less economically burdensome. Passing of the Obama-era ACA was a positive step forward as evident by the decline in uninsured U.S. citizens (estimated 7–16.4 million) and Medicare's lower rate of spending following the legislation [43]. The U.S. must resist the current political efforts to dislodge the inclusive tenets of the Affordable Care Act. Again, this is not to suggest that universal healthcare will be a cure-all, as social determinants of health must also be addressed. However, addressing these determinants will take time and universal healthcare for all U.S. citizens is needed now. Only through universal and inclusive healthcare will we be able to pave an economically sustainable path towards true public health.

Author Contributions: Conceptualization, G.Z., Z.Y.K., J.B.M., and L.S.; writing-original draft preparation, G.Z.; writing-review and editing, Z.Y.K., J.B.M., and L.S.; supervision, L.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. *NBC News/Wall Street Journal Survey*; Study #19175; Hart Research Associates/Public Opinion Strategies: New York, NY, USA, 2019.
2. WHO. Universal Health Coverage. Available online: https://www.who.int/healthsystems/universal_health_coverage/en/ (accessed on 2 October 2020).
3. Light, D.W. Universal health care: Lessons from the British experience. *Am. J. Public Health* **2003**, *93*, 25–30. [CrossRef] [PubMed]
4. Chernichovsky, D.; Leibowitz, A.A. Integrating public health and personal care in a reformed US health care system. *Am. J. Public Health* **2010**, *100*, 205–211. [CrossRef] [PubMed]
5. Unger, J.P.; De Paepe, P. Commercial Health Care Financing: The Cause of U.S., Dutch, and Swiss Health Systems Inefficiency? *Int. J. Health Serv.* **2019**, *49*, 431–456. [CrossRef] [PubMed]
6. Country Rankings: World & Global Economy Rankings on Economic Freedom. Available online: https://www.heritage.org/index/ranking (accessed on 29 September 2020).
7. Economic Freedom of the World. Cato Institute. Available online: https://www.cato.org/economic-freedom-world (accessed on 29 September 2020).
8. Disparities|Healthy People 2020. Available online: https://www.healthypeople.gov/2020/about/foundation-health-measures/Disparities (accessed on 17 January 2020).
9. Fuchs, V.R. How and why US health care differs from that in other OECD countries. *JAMA J. Am. Med. Assoc.* **2013**, *309*, 33–34. [CrossRef] [PubMed]
10. Blahous, C.; Fish, J.; Smith, L.F.; Strategist, S.R. The Costs of a National Single-Payer Healthcare System. 2018. Available online: https://www.mercatus.org/publications/federal-fiscal-policy/costs (accessed on 8 January 2020).
11. Daniels, M.; Panetta, L.; Penny, T. *Primary Care: Estimating Democratic Candidates' Health Plans*; Committee for a Responsible Federal Budget: Washington, DC, USA, 2020.
12. Sessions, S.Y.; Lee, P.R. A road map for universal coverage: Finding a pass through the financial mountains. *J. Health Polit. Policy Law* **2008**, *33*, 155–197. [CrossRef] [PubMed]
13. Sanders, B. Options to Finance Medicare for All. 2020. Available online: https://www.sanders.senate.gov/download/options-to-finance-medicare-for-all?inline=file (accessed on 8 January 2020).
14. Thorpe, K.E. An Analysis of Senator Sanders Single Payer Plan. 2016. Available online: https://berniesanders.com/issues/ (accessed on 8 January 2020).
15. Skocpol, T. *Boomerang: Health Care Reform and the Turn Against Government*; Norton: New York, NY, USA, 1997; pp. 133–143.
16. Podemska-Mikluch, M. Understanding Medicare's Impact on Innovation: A Framework for Policy Reform. 2018. Available online: https://www.mercatus.org/publications (accessed on 2 October 2020).
17. Barua, B. *Waiting Your Turn: Wait Times for Health Care in Canada, 2017 Report*; Fraser Institute: Vancouver, CO, Canada, 2016.
18. *Arguing for Universal HealtH Coverage*; World Health Organization: Geneva, Switzerland, 2013.
19. *A Vision for Primary Care in the 21st Century: Towards Universal Health Coverage and the Sustainable Development Goals*; UNICEF, World Health Organization: Geneva, Switzerland, 2020.
20. Crowley, R.; Daniel, H.; Cooney, T.G.; Engel, L.S. Envisioning a better U.S. health care system for all: Coverage and cost of care. *Ann. Intern. Med.* **2020**, *172*, S7–S32. [CrossRef]
21. Doherty, R.; Cooney, T.G.; Mire, R.D.; Engel, L.S.; Goldman, J.M. Envisioning a Better U.S. Health Care System for All: A Call to Action by the American College of Physicians. *Ann. Intern. Med.* **2020**, *172*, S3. [CrossRef]

22. Ward, B.W.; Schiller, J.S.; Goodman, R.A. Multiple chronic conditions among US adults: A 2012 update. *Prev. Chronic Dis.* **2014**, *11*, E62. [CrossRef]
23. Fryar, C.D.; Chen, T.-C.; Li, X. Prevalence of uncontrolled risk factors for cardiovascular disease: United States, 1999–2010. *NCHS Data Brief* **2012**, 1–8. Available online: http://www.ncbi.nlm.nih.gov/pubmed/23101933 (accessed on 13 March 2020).
24. Pampel, F.C.; Krueger, P.M.; Denney, J.T. Socioeconomic Disparities in Health Behaviors. *Annu. Rev. Sociol.* **2010**, *36*, 349–370. [CrossRef]
25. U.S. Department of Health & Human Services. *Office of the Assistant Secretary for Planning and Evaluation Value of Health Insurance: Few of the Uninsured Have Adequate Resources to Pay Potential Hospital Bills|ASPE*. Available online: https://aspe.hhs.gov/basic-report/value-health-insurance-few-uninsured-have-adequate-resources-pay-potential-hospital-bills (accessed on 13 March 2020).
26. McBride, T.D. Uninsured spells of the poor: Prevalence and duration. *Health Care Financ. Rev.* **1997**, *19*, 145–160.
27. Saydah, S.H.; Imperatore, G.; Beckles, G.L. Socioeconomic status and mortality: Contribution of health care access and psychological distress among U.S. adults with diagnosed diabetes. *Diabetes Care* **2013**, *36*, 49–55. [CrossRef]
28. American Diabetes Association. Economic Costs of Diabetes in the U.S. in 2012. *Diabetes Care* **2013**, *36*, 1033–1046. [CrossRef] [PubMed]
29. Lim, S.S.; Vos, T.; Flaxman, A.D.; Danaei, G.; Shibuya, K.; Adair-Rohani, H.; Amann, M.; Anderson, H.R.; Andrews, K.G.; Aryee, M.; et al. A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990–2010: A systematic analysis for the Global Burden of Disease Study 2010. *Lancet* **2012**, *380*, 2224–2260. [CrossRef]
30. Leng, B.; Jin, Y.; Li, G.; Chen, L.; Jin, N. Socioeconomic status and hypertension: A meta-analysis. *J. Hypertens.* **2015**. [CrossRef]
31. Kirkland, E.B.; Heincelman, M.; Bishu, K.G.; Schumann, S.O.; Schreiner, A.; Axon, R.N.; Mauldin, P.D.; Moran, W.P. Trends in healthcare expenditures among US adults with hypertension: National estimates, 2003–2014. *J. Am. Heart Assoc.* **2018**. [CrossRef] [PubMed]
32. Bhurosy, T.; Jeewon, R. Overweight and obesity epidemic in developing countries: A problem with diet, physical activity, or socioeconomic status? *Sci. World J.* **2014**, *2014*, 964236. [CrossRef]
33. Hammond, R.A.; Levine, R. The economic impact of obesity in the United States. *Diabetes Metab. Syndr. Obes.* **2010**, *3*, 285–295. [CrossRef] [PubMed]
34. Clark, G.; O'Dwyer, M.; Chapman, Y.; Rolfe, B. An Ounce of Prevention is Worth a Pound of Cure-Universal Health Coverage to Strengthen Health Security. *Asia Pac. Policy Stud.* **2018**, *5*, 155–164. [CrossRef]
35. A Better-Quality Alternative: Single-Payer National Health System Reform|Physicians for a National Health Program. Available online: https://www.pnhp.org/publications/a_better_quality_alternative.php?page=all (accessed on 7 October 2020).
36. Fielding, J.E.; Teutsch, S.; Koh, H. Health reform and healthy people initiative. *Am. J. Public Health* **2012**, *102*, 30–33. [CrossRef]
37. CDC The Power of Prevention. 2009. Available online: https://www.cdc.gov/chronicdisease/pdf/2009-power-of-prevention.pdf (accessed on 11 May 2020).
38. Levi, J.; Segal, L.; Juliano, C. *Issue Report: Prevention for a Healthier America: Investments in Disease Prevention Yield Significant Savings, Stronger Communities*; Trust for America's Health: Washington, DC, USA, 2009.
39. Lee, B.Y.; Adam, A.; Zenkov, E.; Hertenstein, D.; Ferguson, M.C.; Wang, P.I.; Wong, M.S.; Wedlock, P.; Nyathi, S.; Gittelsohn, J.; et al. Modeling the economic and health impact of increasing children's physical activity in the United States. *Health Aff.* **2017**. [CrossRef]
40. Hamasaki, H. Daily physical activity and type 2 diabetes: A review. *World J. Diabetes* **2016**. [CrossRef] [PubMed]
41. Magnussen, J.; Vrangbaek, K.; Saltman, R.B. Nordic Health Care Systems Recent Reforms and Current Policy Challenges Nordic Health Care Systems. *Eur. Obs. Heal. Syst. Policies Ser.* **2009**. [CrossRef]

42. Fullman, N.; Yearwood, J.; Abay, S.M.; Abbafati, C.; Abd-Allah, F.; Abdela, J.; Abdelalim, A.; Abebe, Z.; Abebo, T.A.; Aboyans, V.; et al. Measuring performance on the Healthcare Access and Quality Index for 195 countries and territories and selected subnational locations: A systematic analysis from the Global Burden of Disease Study 2016. *Lancet* **2018**. [CrossRef]
43. Blumenthal, D.; Abrams, M.; Nuzum, R. The Affordable Care Act at 5 Years. *N. Engl. J. Med.* **2015**, *372*, 2451–2458. [CrossRef]

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

ELSEVIER

Institut Pasteur

Research in Microbiology 166 (2015) 326–331

www.elsevier.com/locate/resmic

Genomics of *Clostridium tetani*

Holger Brüggemann [a], Elzbieta Brzuszkiewicz [b], Diana Chapeton-Montes [c], Lucile Plourde [d], Denis Speck [d], Michel R. Popoff [c,*]

[a] *Department of Biomedicine, Aarhus University, Denmark*
[b] *Department of Genomics and Applied Microbiology, Institute of Microbiology and Genetics, Georg-August-University of Göttingen, Germany*
[c] *Institut Pasteur, Bactéries anaérobies et Toxines, 28 rue du Dr Roux, F-75724 Paris Cedex 15, France*
[d] *Sanofi-Pasteur, Marcy l'Etoile, France*

Received 22 December 2014; accepted 13 January 2015
Available online 29 January 2015

Abstract

Genomic information about *Clostridium tetani*, the causative agent of the tetanus disease, is scarce. The genome of strain E88, a strain used in vaccine production, was sequenced about 10 years ago. One additional genome (strain 12124569) has recently been released. Here we report three new genomes of *C. tetani* and describe major differences among all five *C. tetani* genomes. They all harbor tetanus-toxin-encoding plasmids that contain highly conserved genes for TeNT (tetanus toxin), TetR (transcriptional regulator of TeNT) and ColT (collagenase), but substantially differ in other plasmid regions. The chromosomes share a large core genome that contains about 85% of all genes of a given chromosome. The non-core chromosome comprises mainly prophage-like genomic regions and genes encoding environmental interaction and defense functions (e.g. surface proteins, restriction-modification systems, toxin-antitoxin systems, CRISPR/Cas systems) and other fitness functions (e.g. transport systems, metabolic activities). This new genome information will help to assess the level of genome plasticity of the species *C. tetani* and provide the basis for detailed comparative studies.

© 2015 Institut Pasteur. Published by Elsevier Masson SAS. All rights reserved.

Keywords: Clostridium tetani; Tetanus neurotoxin; Genomics; Plasmid

1. Introduction

Clostridium tetani is a Gram-positive, spore-forming and anaerobic rod-shaped bacterium which produces a potent neurotoxin, the tetanus toxin (TeNT). *C. tetani* forms spores which appear as translucent terminal enlargements having the general appearance of drumsticks. Spores are resistant to desiccation, heat and exposure to oxygen, and thus *C. tetani* can survive in environments such as soil for long periods of time [1]. More rarely, *C. tetani* can be found in the intestinal flora of animals and humans. Usually, *C. tetani* enters an organism through an open wound. Spores germinate and *C. tetani* grow in anaerobic conditions in necrotic tissues. TeNT is synthesized at the end of exponential growth and is released in the surrounding tissues. TeNT enters motorneurons and is retrogradely transported to the central nervous system. Then, TeNT enters inhibitory interneurons and blocks the release of neurotransmitters (glycine, GABA) upon proteolytic cleavage of the SNARE protein VAMP2 [2,3]. TeNT is responsible for all the symptoms of tetanus, characterized by spastic paralysis (muscle rigidity and spasms) [4]. In contrast to *botulinum* neurotoxins (BoNT), which show wide genetic diversity (see companion articles in this issue), only one TeNT type has been reported until now. It is presumed that *C. tetani* strains exhibit limited genetic variability. A previous study conducted via DNA/DNA hybridization showed that 10 neurotoxigenic and 3 non-toxigenic *C. tetani* strains were homogeneous [5]. At least two genetic profiles have been evidenced

---

* Corresponding author.
E-mail address: mpopoff@pasteur.fr (M.R. Popoff).

http://dx.doi.org/10.1016/j.resmic.2015.01.002
0923-2508/© 2015 Institut Pasteur. Published by Elsevier Masson SAS. All rights reserved.

in five *C. tetani* strains, as tested by pulse field gel electrophoresis [6]. It was first demonstrated that the *tent* gene is located on a large size plasmid and that the loss of this plasmid yields non-toxigenic strains [7]. The first complete genome sequence of a toxigenic *C. tetani* strain (E88) was determined in 2003 [8]. The genome contained a chromosome of approximately 2.8 Mb and the tetanus-toxin-encoding plasmid of 74 kb. One additional *C. tetani* genome was determined in 2014 [9]. Here we report the genome sequences of three additional *C. tetani* strains. We performed comparative genome studies in order to define the core and flexible genome content of *C. tetani*.

2. Materials and methods

2.1. Bacterial strains

The three *C. tetani* strains used for this study were CN655, 184.08 and strain A. Strain CN655 is a collection strain from Prevot's collection (Institut Pasteur) that originated from the Wellcome Institute (1947). Strain 184.08 was isolated from a suppurative arthritis case subsequent to a traumatism in a 75-year-old man in the south of France in 2008. Strain A is a variant of the Harvard strain.

C. tetani strains were grown in TGY broth (pH 7.5) containing trypticase (trypticase-glucose yeast peptone BBL, BD Biosciences; 30 g/L), yeast extract (Bacto Yeast Extract, BD Biosciences; 20 g/L), glucose (5 g/L) and cysteine, HCl (0.5 g/L) under anaerobic conditions.

2.2. DNA extraction and genome sequencing

Genomic DNA from all strains of *C. tetani* was extracted and purified as previously described [10]. Draft sequences of the genomes were obtained by using Illumina/Solexa GAIIx machines at the Beijing Genomics Institute (BGI) (Shenzhen, China). The assembly of sequence reads was done using SOAPdenovo (version 1.05).

2.3. Sequence analysis

Automatic annotation of all contigs was done using RAST (http://rast.nmpdr.org/). To identify homologs in different strains, comparisons of all coding DNA sequences (CDS) were made using a protein sequence-based bidirectional BLAST approach (blastP version 2.2.18). Sequence homologies were only mentioned in this study for proteins with an amino acid identity of >25% and an overlap of the query and subject sequence of >75%. Genome representations were created by a DNA plotter (Sanger Institute).

2.4. Sequence accession

GenBank accession numbers for the draft genome sequences are JSWC00000000 for strain CN655, JSWD00000000 for strain 184.08 and JWIX00000000 for strain A.

3. Results and discussion

Genomes of neurotoxigenic clostridia have been sequenced. Whereas many genome sequences of *C. botulinum* have been determined in the last decade (over 10 completed genomes and more than 30 draft genomes), only two *C. tetani* genomes are available thus far. The first genome sequence of *C. tetani* was determined in 2003 [8]. Recently, the genome of strain 12124569 was sequenced, an isolate that caused a rare case of chronic osteitis [9]. We decided to obtain more information on the genomic diversity of *C. tetani* by sequencing three additional genomes, those of strains CN655, 184.08 and A. Strain 184.08 was isolated from a suppurative arthritis of a 75-year-old man in the south of France in 2008. CN655 is a collection strain from Prevot's collection (Institut Pasteur) that originated from the Wellcome Institute (1947). Strain A is a variant of a Harvard strain. All these strains produce the tetanus toxin as tested in a mouse bioassay (data not shown).

3.1. Tetanus-toxin-encoding plasmids

All five strains possess a tetanus-toxin-encoding plasmid. G + C content, size and number of coding sequences (CDS) vary slightly, with the exception of strain 12124569 that carries a considerably smaller plasmid (Table 1). Sequence comparison revealed that plasmids of strains E88, CN665 and A (hereafter named pE88, pCN665 and pA) are highly similar to each other (Fig. 1); in contrast, plasmids of strains 184.08 and 12124569 (named p184.08 and p12124569, respectively) strongly differ (Fig. 2). Bidirectional blast (BIBlast) comparison revealed the extent of sequence divergence (Supplemental Table 1). Several genes of pE88, pCN655 and pA are not present on p184.08 and p12124569, including a two-component system and an ABC transport system. In contrast, p184.08 contains additional genes, including a type III restriction-modification system, a phage integrase and a putative bacitracin resistance transport system (Supplemental table 1B).

The sequences of TeNT and collagenase ColT are highly conserved: TeNT of pE88, pCN655 and pA are 100% identical at the amino acid level and are 98% identical (1294/1315) to TeNT from strains 184.08 and 12124569. Similarly, ColT of pE88, pCN655 and pA are 100% identical and they are 97% (963/991) identical to ColT of p184.08. Other highly conserved proteins encoded by all five plasmids are the two

Table 1
Features of tetanus-toxin-encoding plasmids.

Plasmid name	Size	G + C (%)	CDS[b]
pE88	74.0 kb	24.5	98
p184.08	ca. 73.0 kb[a]	24.9	93
pA	ca. 71.7 kb[a]	24.5	90
pCN655	ca. 69.6 kb[a]	24.3	90
p12124569	58.4 kb	24.0	77

[a] Multiple contigs.

[b] Based on RAST annotation (http://rast.nmpdr.org/).

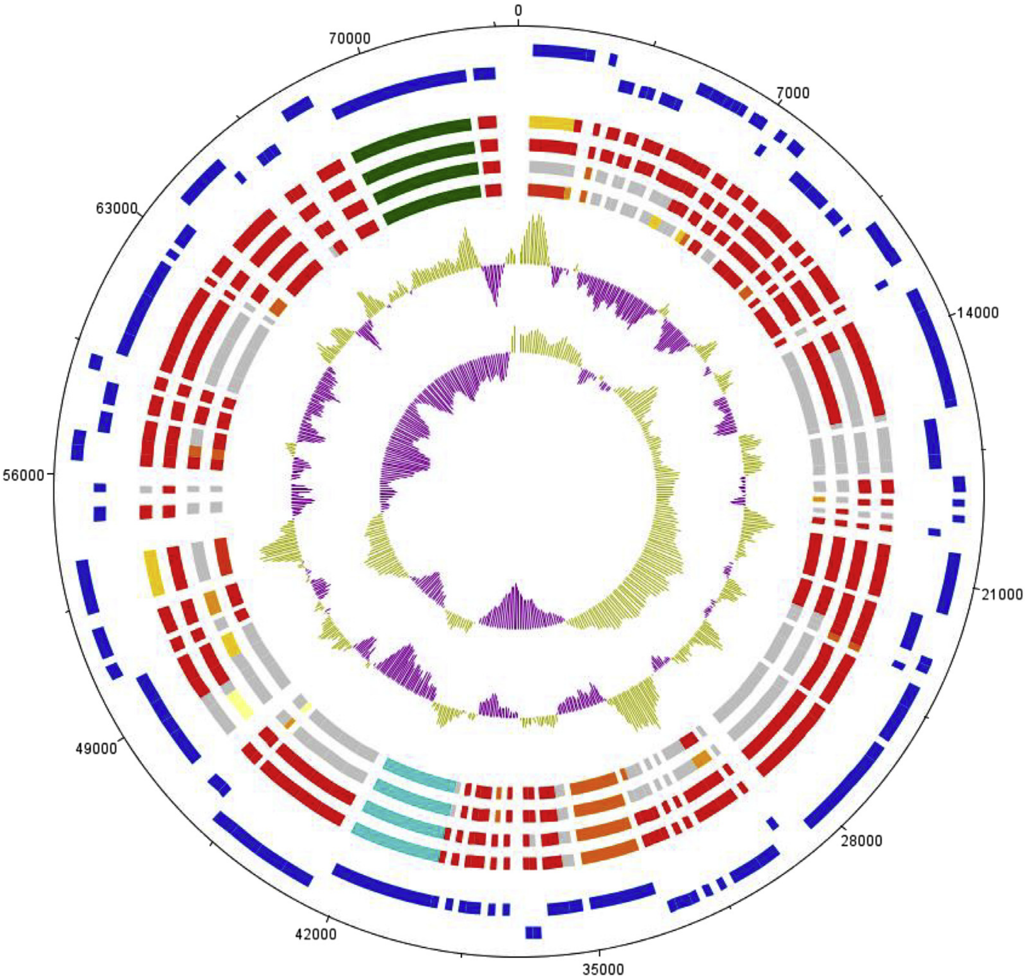


Fig. 1. A BiBlast-based comparison of five tetanus-toxin-encoding plasmids of *C. tetani* with pE88 as reference plasmid. The blue outer rings depict the CDS of pE88 (plus and minus strand); the four inner rings depict bidirectional Blast results regarding pCN655, pA, p12124569 and p184.08, respectively. Red: high similarity; orange: medium similarity; yellow: low similarity; gray: absence. The highly conserved tetanus toxin and collagenase genes are depicted in dark green and cyan, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.) two inner rings depict the G + C content and the G + C skew.

putative RNA polymerase sigma factors (CTP10 and CTP11 in pE88; 94% and 99% protein identity, respectively), an ABC transport system (CTP30/CTP31) and two putative aminopeptidases (CTP20 and CTP56 in pE88; 97% and 98% protein identity, respectively).

3.2. Chromosome of C. tetani

The chromosomes of the five *C. tetani* strains have a similar size (ca. 2.8 Mb) and G + C content (28.6–28.8%) (Table 2). They are highly conserved: the core genome comprises 85% and 83% of all CDS in strains E88 and 184.08, respectively, when applying a cutoff of 30% protein identity (Fig. 3). Even with a cutoff of 50% protein identity, 84% and 81.5% of all CDS of strains E88 and 184.08, respectively, belong to the core genome (Supplemental Table 2).

The majority of strain-specific functions are encoded in prophage regions. For example, the chromosome of strain E88 contains three strain-specific prophage regions (Supplemental Table 2). Other strain-specific functions include transposases (in particular, in the genome of E88), cell-surface/adhesion/S-layer proteins, restriction-modification-systems, ABC transport systems and efflux pumps, two-component systems, resistance determinants and defense/repair functions as well as specific metabolic functions. For instance, strains 184.08 and 12124569 contain peptide methionine sulfoxide reductase MsrB. Strain E88 specifically contains a plasmid-like locus comprising an adenine-specific methyltransferase and a toxin-antitoxin system. Regarding metabolic functions, for example, strain 184.08 possesses the genes for an allophanate hydrolase that converts allophanate to ammonium and carbon dioxide, an essential system for using urea as a nitrogen source.

Another difference between the strains concerns the clustered regularly interspaced short palindromic repeats (CRISPR)/Cas loci and the diversity of CRISPR spacers. The CRISPR/Cas system constitutes an adaptive immune system that protects the bacterium from bacteriophage infections and from mobile genetic elements. Strain E88 contains two loci:

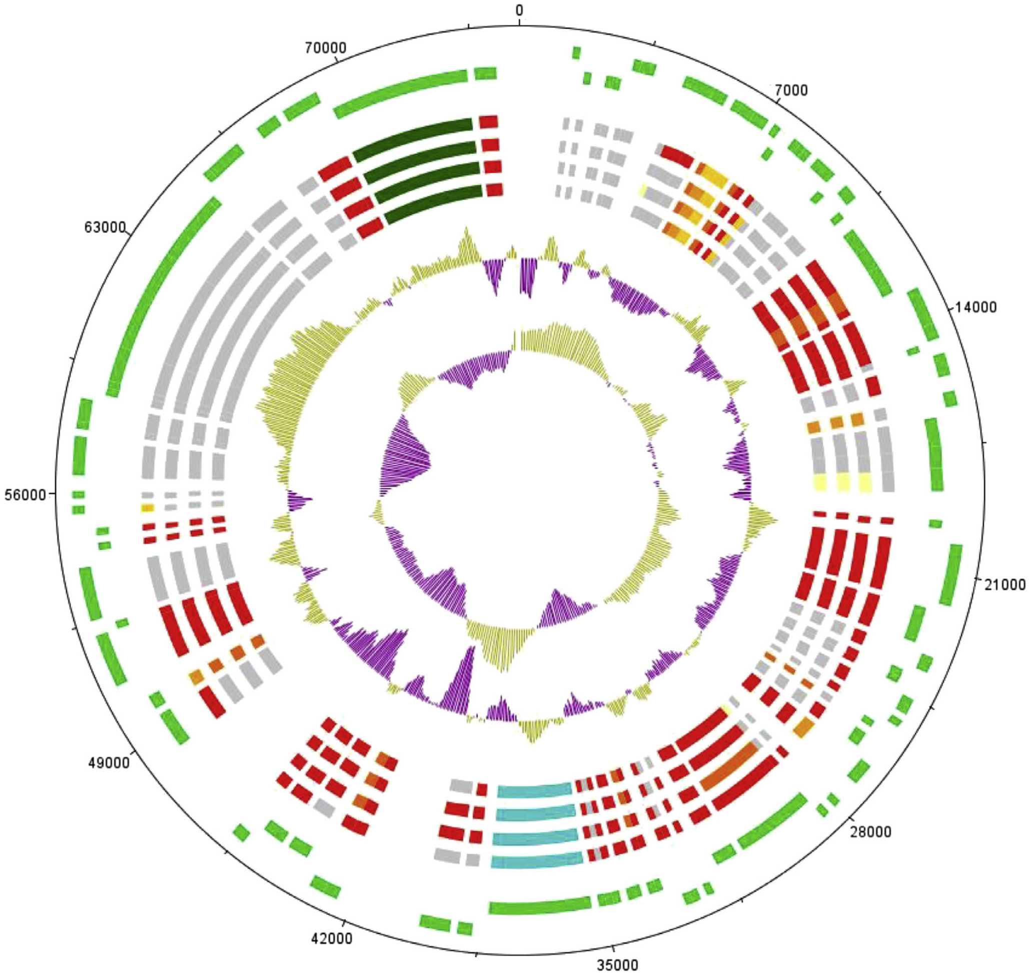


Fig. 2. A BiBlast-based comparison of five tetanus-toxin-encoding plasmids of *C. tetani* with p184.08 as reference plasmid. The green outer rings depict the CDS of p184.08 (plus and minus strand); the four inner rings depict BiBlast results regarding p12124569, pE88, pCN655 and pA, respectively. Red: high similarity; orange: medium similarity; yellow: low similarity; gray: absence. The highly conserved tetanus toxin and collagenase genes are depicted in dark green and cyan, respectively. The two inner rings depict the G + C content and the G + C skew. Please note that the map of p184.08 was reconstructed from 9 contigs aligned to pE88; thus, sequence gaps in p184.08 exist. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

one locus (CTC1142-CTC1148) is composed of the *cas* genes *cas6, csx8 (cas8a1), devR (cas7), cas5, cas3, cas4 and cas1* (type I-A-like system) with a repeat consensus 5′-GTATTAGTAGCACCATATTGGAATGTAAAT-3′ and the other one (CTC1463-CTC1469) is composed of *cas6, csh1 (cas8b), csh2 (cas7), cas5, cas3, cas4, cas1* and *cas2* (type I–B system) with a repeat consensus 5′-ATTTAAATACAACTCTTGTTATTGTTCAAC-3′. Most of the spacers (66 in strain E88) show no similarity to sequences deposited in databases. Some spacers match strain-specific prophage regions in other *C. tetani* and/or *C. botulinum* genomes. Interestingly, a number of spacers also match regions within the same genome: for example, spacers in E88 match E88 genomic regions encoding stage IV sporulation protein (CTC1140), sporulation sigma-E-factor-processing peptidase (CTC1125), vancomycin b-type resistance protein VanW (CTC0360) and S-layer protein/N-acetylmuramoyl-L-alanine amidase (CTC0462). This indicates that such regions are located on mobile genetic elements that have been acquired by horizontal gene transfer events. The type IA-like CRISPR/Cas system in strain 12124569 is associated with a different repeat (5′-AATTTACATTCCAATATGGTGCTACTAATAC-3′) and contains 22 spacers. Some of them match prophage-like genomic regions in *C. tetani* and *C. botulinum*. The genome of strain 184.08 does not contain any CRISPR/Cas region. Interestingly, this genome also contains more prophage-like genomic regions that account for its larger genome size

Table 2
Features of 5 chromosomes of *C. tetani* strains.

	Size	Contigs	G + C (%)	CDS[a]
E88	2799251	1	28.7	2736
CN655	ca. 2780622	111	28.6	2734
A	ca. 2754351	90	28.6	2704
12,124,569	2807472	1	28.8	2809
184.08	ca. 2841472	135	28.6	2826

[a] Based on RAST annotation (http://rast.nmpdr.org/).

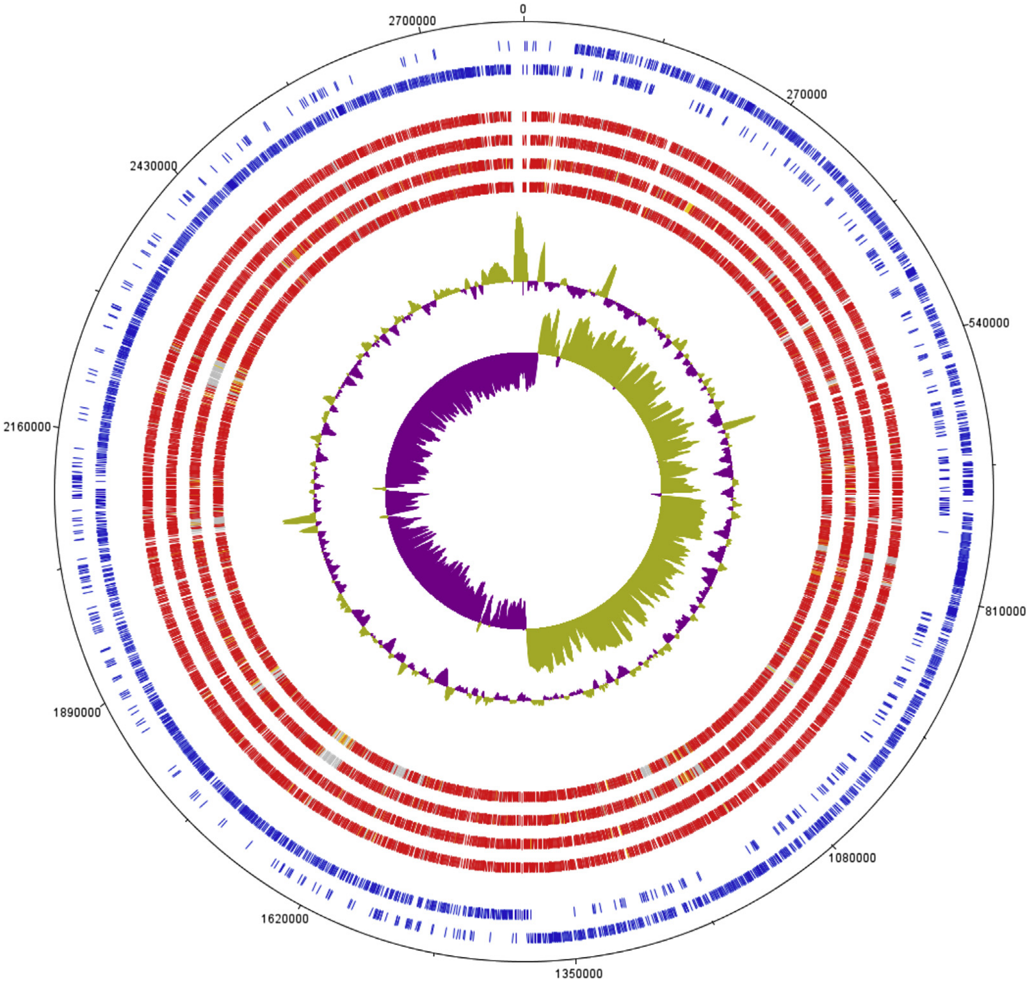


Fig. 3. A bidirectional blast comparison of the chromosomes of five *C. tetani* strains. The outer blue ring depict the CDS of reference strain E88; the inner four rings depict BiBlast results in the genomes of CN655, A, 12124569 and 184.08, respectively. Red: high similarity, orange: medium similarity, yellow: low similarity, gray: absence. The two inner rings depict G + C content and the G + C skew. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

compared to the other four *C. tetani* genomes (Table 2; Supplemental Table 2B). This indicates that the lack of CRISPR/Cas systems renders strain 184.08 more vulnerable to phage infections. The divergence of the CRISPR/Cas regions of *C. tetani* reflects the distinct bacteriophage infection history of each individual *C. tetani* strain.

In conclusion, genome comparison of *C. tetani* highlights flexible, non-core genome content. Larger islands within individual genomes often constitute prophage-like regions. Smaller genomic regions of divergence encode a set of fitness functions, increase competitiveness and protection and guarantee survival under stress conditions and attack, e.g. against bacteriophage infections.

The tetanus-toxin-encoding plasmids of the five isolates have several differences, underlying their highly flexible nature compared to the chromosome. In particular, plasmids of strains 184.08 and 12124569 differ significantly from pE88, the plasmid from the vaccine-producing strains E88 and A.

Conflict of interest

The authors have no conflict of interest.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.resmic.2015.01.002.

References

[1] Popoff MR, Mazuet C, Poulain B. Botulism and tetanus. The prokaryotes: human microbiology. 4th edn. Berlin Heidelberg: Springer-Verlag; 2013. p. 247–90.

[2] Bohnert S, Schiavo G. Tetanus toxin is transported in a novel neuronal compartment characterized by a specialized pH regulation. J Biol Chem 2005;280:42336–44.

[3] Rossetto O, Scorzeto M, Megighian A, Montecucco C. Tetanus neurotoxin. Toxicon 2013;66:59–63. http://dx.doi.org/10.1016/j.toxicon.2012.1012.1027 [Epub 2013 Feb 1016].

[4] Hassel B. Tetanus: pathophysiology, treatment and the possibility of using botulinum toxin against tetanus-induced rigidity and spasms. Toxins Basel 2013;5:73−83. http://dx.doi.org/10.3390/toxins5010073.

[5] Nakamura S, Okado I, Abe T, Nishida S. Taxonomy of *Clostridium tetani* and related species. J Gen Microbiol 1979;113:29−35.

[6] Plourde-Owobi L, Seguin D, Baudin MA, Moste C, Rokbi B. Molecular characterization of *Clostridium tetani* strains by pulsed-field gel electrophoresis and colony PCR. Appl Environ Microbiol 2005;71:5604−6.

[7] Finn CW, Silver RP, Habig WH, Hardegree MC, Zon G, Garon CF. The structural gene for tetanus neurotoxin is on a plasmid. Science 1984;224:881−4.

[8] Brüggemann H, Bäumer S, Fricke WF, Wiezr A, Liesagang H, Decker I, et al. The genome sequence of *Clostridium tetani*, the causative agent of tetanus disease. Proc Ntl Acad Sci USA 2003;100:1316−21.

[9] Levy PY, Fournier PE, Lotte L, Million M, Brouqui P, Raoult D. *Clostridium tetani* osteitis without tetanus. Emerg Infect Dis 2014;20:1571−3. http://dx.doi.org/10.3201/eid2009.131579.

[10] Popoff MR, Guillou JP, Carlier JP. Taxonomic position of lecithinase-negative strains of *Clostridium sordellii*. J Gen Microbiol 1985;131:1697−703.

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/366731938

Investigation of susceptibility genes for chickenpox disease across multiple continents

Article *in* Biochemistry and Biophysics Reports · March 2023

DOI: 10.1016/j.bbrep.2022.101419

CITATIONS
11

READS
142

11 authors, including:

Lalu Muhammad Irham
Universitas Ahmad Dahlan

115 PUBLICATIONS **7,991** CITATIONS

SEE PROFILE

Wirawan Adikusuma
National Research and Innovation Agency

83 PUBLICATIONS **2,217** CITATIONS

SEE PROFILE

Lolita Lolita
Nanjing Medical University

26 PUBLICATIONS **79** CITATIONS

SEE PROFILE

Made Ary Sarasmita
Udayana University

64 PUBLICATIONS **11,282** CITATIONS

SEE PROFILE

All content following this page was uploaded by Gina Noor Djalilah on 04 January 2023.

The user has requested enhancement of the downloaded file.

Biochemistry and Biophysics Reports 33 (2023) 101419

Contents lists available at ScienceDirect

Biochemistry and Biophysics Reports

journal homepage: www.elsevier.com/locate/bbrep

Investigation of susceptibility genes for chickenpox disease across multiple continents

Lalu Muhammad Irham [a,*], Wirawan Adikusuma [b], Lolita Lolita [a], Anisa Nova Puspitaningrum [a], Arief Rahman Afief [a], Made Ary Sarasmita [c], Haafizah Dania [a], Sabiah Khairi [d], Gina Noor Djalilah [e], Barkah Djaka Purwanto [f,g], Rocky Cheung [h]

[a] *Faculty of Pharmacy, Universitas Ahmad Dahlan, Yogyakarta, Indonesia*
[b] *Department of Pharmacy, University of Muhammadiyah Mataram, Mataram, Indonesia*
[c] *Pharmacy Study Program, Faculty of Science and Mathematics, Udayana University, Bali, Indonesia*
[d] *School of Nursing, College of Nursing, Taipei Medical University, Taipei, 11031, Taiwan*
[e] *Medical Faculty Muhammadiyah Surabaya, Surabaya, 60115, Indonesia*
[f] *Faculty of Medicine, University of Ahmad Dahlan, Yogyakarta, 55191, Indonesia*
[g] *PKU Muhammadiyah Bantul Hospital, Bantul, Yogyakarta, 55711, Indonesia*
[h] *Department of Chemistry and Biochemistry, University of California, Los Angeles, USA*

ARTICLE INFO

Keywords:
Chickenpox
Genomic variants
GWAS
HLA-S
HCG4P5
ABHD16A

ABSTRACT

Chickenpox (varicella) is caused by infection with the varicella-zoster virus (VZV), a neurotropic alpha herpes virus with a double-stranded DNA genome. Chickenpox can cause life-threatening complications, including subsequent bacterial infections, central nervous system symptoms, and even death without any risk factors. Few studies have been reported to investigate genetic susceptibility implicated in chickenpox. Herein, our study identified global genetic variants that potentially contributed to chickenpox susceptibility by utilizing the established bioinformatic-based approach. We integrated several databases, such as genome-wide association studies (GWAS) catalog, GTEx portal, HaploReg version 4.1, and Ensembl databases analyses to investigate susceptibility genes associated with chickenpox. Notably, increased expression of *HLA-S, HCG4P5*, and *ABHD16A* genes underlie enhanced chickenpox susceptibility in the European, American, and African populations. As compared to the Asian population, Europeans, Americans, and Africans have higher allele frequencies of the extant variants rs9266089, rs10947050, and rs79501286 from the susceptibility genes. Our study suggested that these susceptibility genes and associated genetic variants might play a critical role in chickenpox progression based on host genetics with clinical implications.

1. Introduction

Chickenpox, also known as varicella, is an infectious disease caused by the varicella-zoster virus (VZV) [1]. The VZV, also known as herpesvirus 3, is a type of alpha herpesvirus which has double-stranded Deoxyribonucleic Acid (DNA). The primary targets of VZV infection are T lymphocytes, ganglia, and epithelial cells [2]. The most common symptoms of chickenpox in children are fever, headache, backache, chills, nausea, and vomiting [3]. These symptoms are often accompanied by the appearance of a rash that spreads across the entire body's surface skin [4].

VZV infections occur globally, but the epidemiology of these infections varies by region depending on the geographical location. This infection typically occurs in tropical and subtropical regions. In temperate zones, the incidence of chickenpox infection is quite high during the winter and early summer [5]. The majority of individuals living in the United States and Great Britain have had chickenpox by age 15 [6]. In other temperate regions, such as Slovenia, three-to four-year-old children account for 75% of chickenpox infections [7]. In tropical and subtropical climates, adults are more frequently infected with varicella than children [8].

VZV DNA detection in blood and saliva are helpful for the diagnosis and prognosis of varicella/chickenpox infections. VZV DNA can be found in T cells after 10 days of rash and remains there a week after [9].

---

* Corresponding author.
E-mail address: lalu.irham@pharm.uad.ac.id (L.M. Irham).

https://doi.org/10.1016/j.bbrep.2022.101419
Received 4 October 2022; Received in revised form 1 December 2022; Accepted 23 December 2022
2405-5808/© 2022 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

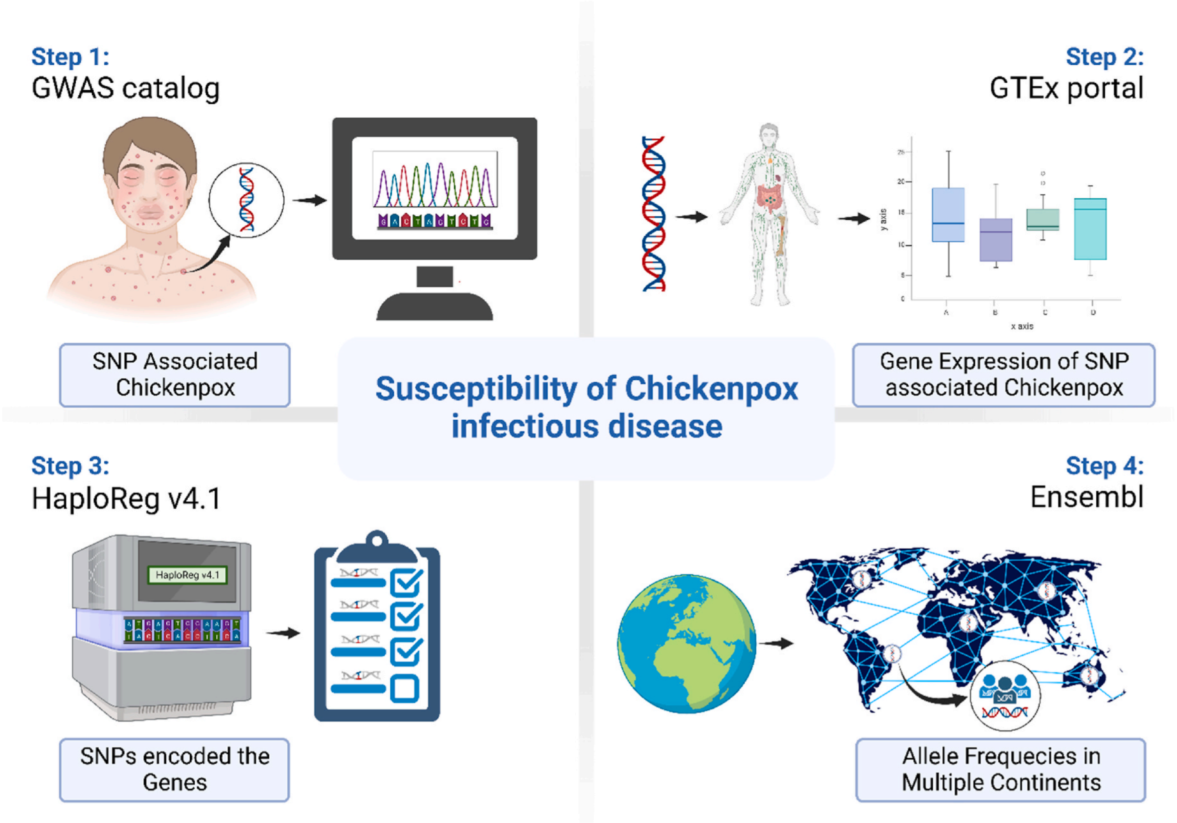

Step 1:
GWAS catalog
SNP Associated Chickenpox
Step 2:
GTEx portal
y axis
x axis
A
B
C
D
Gene Expression of SNP associated Chickenpox
Susceptibility of Chickenpox infectious disease
Step 3:
HaploReg v4.1
HaploReg v4.1
SNPs encoded the Genes
Step 4:
Ensembl
Allele Frequecies in Multiple Continents

Fig. 1. Schematic of custom bioinformatics pipeline to identify chickenpox disease susceptibility genes across multiple continents. This figure was created with BioRender.com under agreement number "BH24OEAKNY ".

Variations in genes can aggravate chickenpox infection, but it is questionable whether they affect gene expression. Furthermore, a population with a high proportion of susceptibility gene carriers should be more susceptible to the disease.

To date, few studies have linked host genetic variation to chickenpox susceptibility [10]. As of this time, a previous study primarily focuses on the identification of the VZV genotype. The goal of the current study is to identify genetic variations that may contribute to Chickenpox disease transmission by utilizing bioinformatic-based approach. We integrated the genome-wide association studies (GWAS) catalog database with genomic datasets to identify gene expression associated with chickenpox. Therefore, our findings will provide the opportunity for future research to investigate whether these variants might be associated with different risks and susceptibility to chickenpox infection. Additionally, this discovery can be used to identify new candidate drugs for treating chickenpox disease.

2. Methods

A summary of the methodology of screening variants-associated with chickenpox disease is illustrated across the various steps in the bioinformatics pipeline (Fig. 1). We leveraged several bioinformatic-based approaches to integrate the variants of chickenpox disease, including GWAS catalog, GTEx portal, HaploReg version 4.1, and Ensembl databases. In the present study, the initial screening of Single Nucleotide Polymorphism (SNP) associated with chickenpox was obtained from the GWAS catalog database [https://www.ebi.ac.uk/gwas/] (accessed on August 19, 2022). We used the keyword "*chickenpox and susceptibility of chickenpox* "to obtain the GWAS-associated chickenpox. To determine the SNP associated with chickenpox, we strictly selected p-value $< 10^{-8}$ as inclusion criteria. In our analysis, we removed all duplicate variants from the dataset in order to guarantee these variants are unique SNPs. Subsequently, a bioinformatic-based approach was employed to interrogate the expression of these variants**.** Next, we validated the SNPs that had been rigorously filtered using the GTEx portal database [http://www.gtexportal.org/home/] (accessed on August 29, 2022). The purpose of this step is to evaluate these genetic variants with gene expression profiles in numerous body tissues. Following a search through the GTEx portal database, we further retrieved the SNPs encoded genes to identify the gene expression profiles. Herein, we leveraged the expression quantitative trait loci (eQTL) are regions harboring nucleotides correlated with alterations in gene expression. Therefore, the variants may cause changes in gene expression in the direction of the tissues involved (i.e., our analyses focused on the skin). Chickenpox is a highly contagious viral infection that causes a blistering, itchy rash on the skin accompanied by fever, viremia, and vesicular lesions that spread over the skin and are frequently reported in children [9]. We used Haploreg version 4.1 and Ensembl genomic databases to map the epidemiological distribution of the genetic basis of allele frequency in different populations. The allele frequencies associated with chickenpox were retrieved from a genetic database such as Haploreg version 4.1 [https://pubs.broadinstitute.org/mammals/haploreg/haploreg.php] (accessed on August 29, 2022) and Ensembl Genome Browser [http://www.ensembl.org/index.html] (accessed on August 30, 2022). Detection of the allele frequencies of each variant is intended to confirm the allele frequencies in populations from multiple continents (Africa, America, Asia, and Europe).

3. Results

Chickenpox-related SNPs were extracted from the GWAS catalog database. We identified 33 SNPs related to chickenpox that passed the inclusion criteria (Supplementary Table 1). Then, SNP data was filtered based on p-value $<10^{-8}$ and three SNPs were obtained related to

Table 1
Chickenpox associated Variant with criterion p-value $<10^{-8}$.

SNP	p-value	chr_pos
rs9266089	$1{,}00 \times 10^{-10}$	31352732
rs10947050	$1{,}00 \times 10^{-10}$	30072514
rs79501286	$2{,}00 \times 10^{-09}$	31475135

chickenpox (Table 1). Using the GTEx portal database, we later evaluated the tissue expression of three SNPs associated with chickenpox. This procedure is intended to verify that the three SNPs are closely associated with gene expression in skin tissues. Our study proposed that the SNPs, including rs9266089, rs10947050, and rs79501286 could affect the tissue expression of chickenpox.

4. Chickenpox gene expression in various tissues

We utilized the GTEx portal database in order to evaluate the expression of related genes in human tissues. This database focuses on the expression of eQTL in various human tissues, especially for the skin. The eQTL analysis could display gene expression data and describes the differences in gene expression caused by the tissue's natural genetic variations [10]. Intriguingly, the GTEx portal database identified three symbol genes associated with chickenpox, namely *HLA-S, HCG4P5*, and *ABHD16A* (Table 2), which exhibit gene expression in the skin. It is important to note that patients with chickenpox typically present with a rash and fever [9]. Therefore, our study focused on gene and variant expression in skin tissues.

5. *Cis*-eQTL of *HLA-S, HCG4P5,* and *ABHD16A genes* in skin tissue

In the present study, we utilized the GTEx portal database with the knowledge that the functional roles of these variants affect protein expression. The eQTL associated with chickenpox expression in skin tissue were identified using the GTEx portal database. Forty different gene variants demonstrated the expression in skin tissues. However, only three genes respectively demonstrated a dominant effect on the expression of *HLA-S* (Fig. 2A), *HCG4P5* (Fig. 2B), and *ABHD16A* (Fig. 2C) in the skin. Furthermore, an additional search using PubMed database text mining was conducted to investigate the three associated eQTL in the skin tissues to determine whether the genetic variants might have clinical significance.

6. Relationship between gene variation and gene expression in skin tissue

Using the publicly available GTEx portal database, we identified minor alleles of *HLA-S* rs9266089, *HCG4P5* rs10947050, and *ABHD16A* rs79501286 respectively associated with chickenpox. As shown in Table 2 and Fig. 3, the *HLA-S* rs9266089 AA genotype was associated with increased expression in skin tissue compared to the GA and GG genotypes. The *HCG4P5* rs10947050 TT genotype had a greater expression in the skin than the CT and CC genotypes. Meanwhile, the *ABHD16A* rs79501286 AA genotype had a higher expression in the skin than the GA and GG genotypes.

7. Allele frequencies of selected variants in different populations

The genetic basis of allele frequency of three genes *HLA-S*, *HCG4P5,* and *ABHD16A* expression and function in different populations is still largely unknown. Therefore, we determined allele frequencies from human populations across multiple continents after identifying chickenpox expression variations. Table 3 shows allele frequencies in African, American, Asian, and European populations. The above allele frequencies were obtained from the Haploreg version 4.1 database, available at https://pubs.broadinstitute.org/mammals/haploreg/haploreg.php (accessed on August 29, 2022). There was a variation in the allele frequencies across the population for each of the SNPs shown in Fig. 4, which were obtained from the Ensembl Genome Browser database http://www.ensembl.org/Homo_sapiens/Variation (accessed on August 30, 2022). Table 3 and Fig. 4 reveal the global allele frequencies for each SNP. The frequency of the A allele at rs9266089, which was associated with the highest expression of *HLA-S* in chickenpox, was significantly lower in Asian populations (2%) than in American (11%), African (10%), or European (15%) populations. T allele frequency showed a higher expression in *HCG4P5* rs10947050, whereas it showed much a lower expression in African (13%) populations compared to Asians (28%), Americans (23%), and Europeans (31%). The A allele of rs79501286 in *ABHD16A* was much less common in Africa (1%), Asia (1%), and America (4%), than in Europe (10%). The allele frequencies of *HLA-S* variant rs9266089, *HCG4P5* variant rs10947050, and *ABHD16A* variant rs79501286 indicated that Asian populations might express these genomic variants less than European and American populations. Our current finding suggested that these susceptibility genes and associated genetic variants might play a critical role in chickenpox progression based on host genetics with clinical implications.

8. Discussion

Chickenpox is a virus-indused disease that infects humans and is transmissible via airborne or direct contact [11]. In this study, we examined the influence of *HLA-S*, *HCG4P5*, and *ABHD16A* on the tissue expression of chickenpox. Chickenpox symptoms include fever, headache, backache, chills, nausea, and vomiting [3] and accompanied by a rash on the entire body's skin surface [4]. The gene responsible for chickenpox, *HLA-S*, has been reported as a potential factor in herpes zoster disease [12], whose primary cause is VZV [13]. It is important to note that *HLA* belongs to the human leukocyte antigen (HLA) family, which produces a protein that plays a critical role in the immune system. *HLA-S* is part of a family of genes called *HLA* complex. The previous study reported that the *HLA* complexes assist the immune system in distinguishing between proteins made by the body and those made by foreign invaders such as viruses and bacteria [14]. It is known that the previous study reported the variants of *HLA-A*33* and *HLA-B*44* were at a strongly higher risk of developing Postherpetic neuralgia (PHN) in patients suffering from VZV infectious disease [13]. Other findings supported the study by revealing that the *HLA-A*02* in patients with atherosclerosis is associated with VZV seropositivity [15].

According to HaploReg version 4.1 bioinformatic database revealed that [https://pubs.broadinstitute.org/mammals/haploreg/haploreg.php], these three variants (rs9266089, rs10947050, and rs79501286) were involved in the changing chromatin status of primary T cells from

Table 2
Variants associated with chickenpox were expressed in the skin tissue.

SNP Id	Genecode Id ENSG00000	Gene Symbol	*p*-value	NES	Tissue	Expression
rs9266089	225851.1	*HLA-S*	4.8×10^{-14}	00.54	Skin - Sun Exposed (Lower leg)	AA > GA > GG
rs10947050	227766.1	*HCG4P5*	2.2×10^{-11}	00.36	Skin - Sun Exposed (Lower leg)	TT > CT > CC
rs79501286	204427.11	*ABHD16A*	0.00031	00.18	Skin - Sun Exposed (Lower leg)	AA > GA > GG

The source of expression data was obtained from the GTEx portal database https://www.gtexportal.org/home/

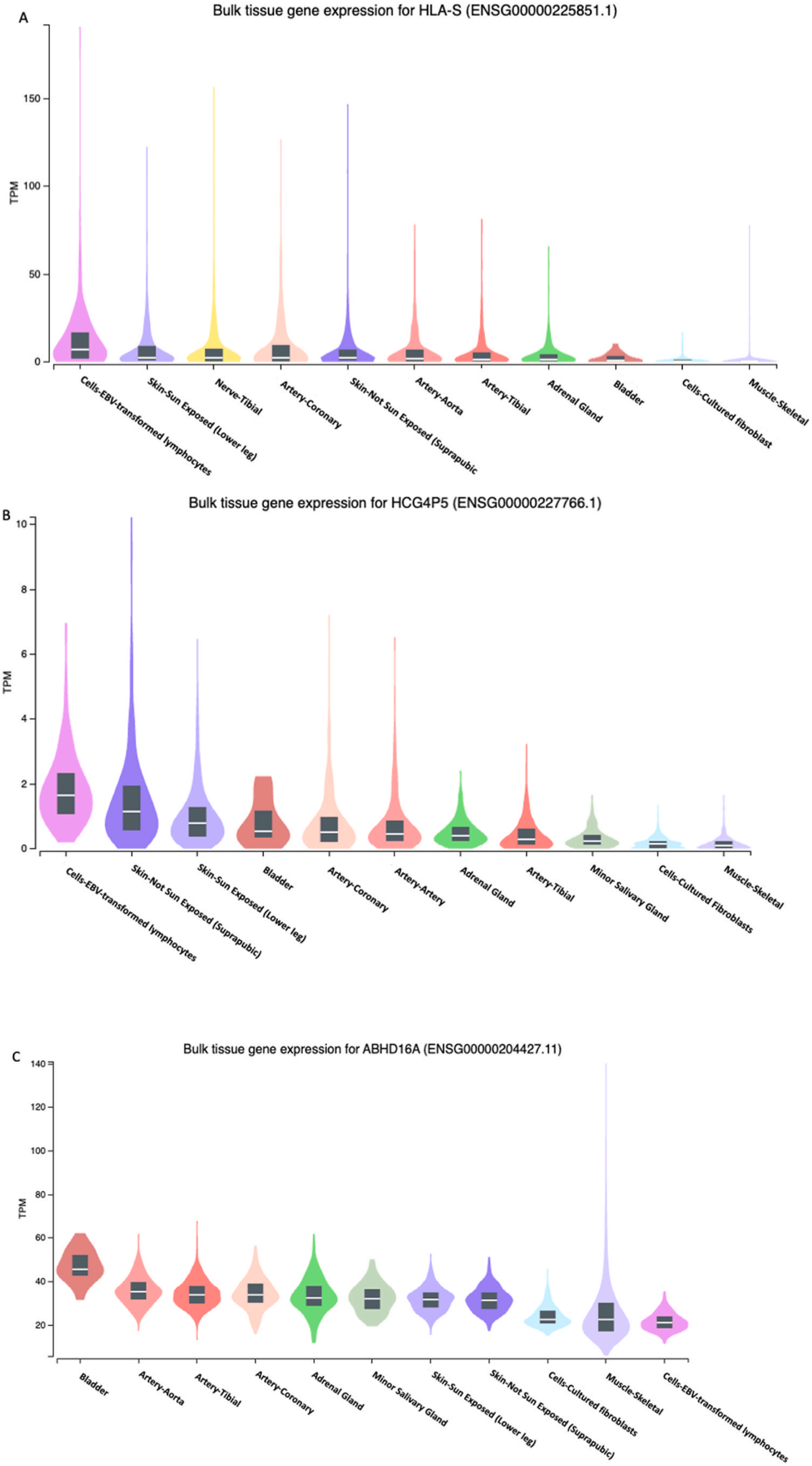


Fig. 2. Expression of chickenpox susceptibility genes across various human tissues: (**A**) *HLA-S*; (**B**) *HCG4P5*, and (**C**) *ABHD16A*.

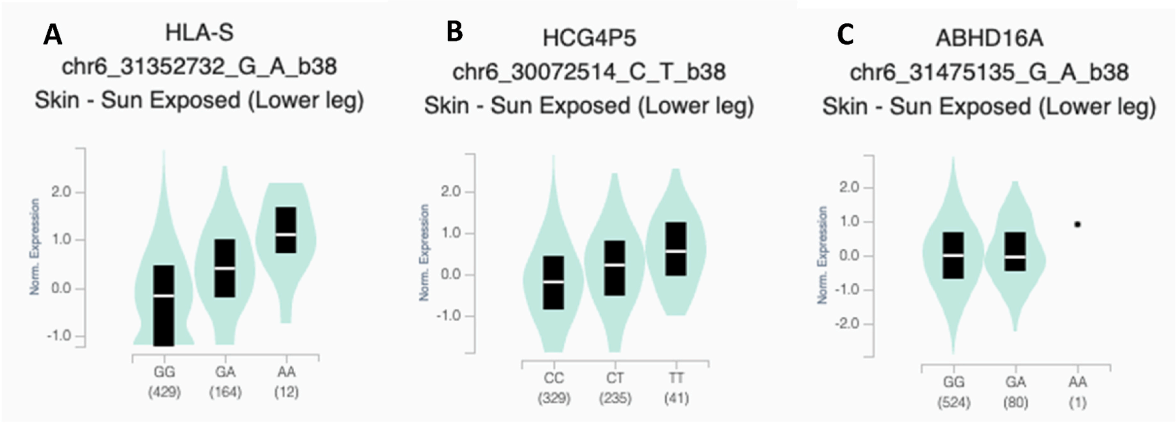


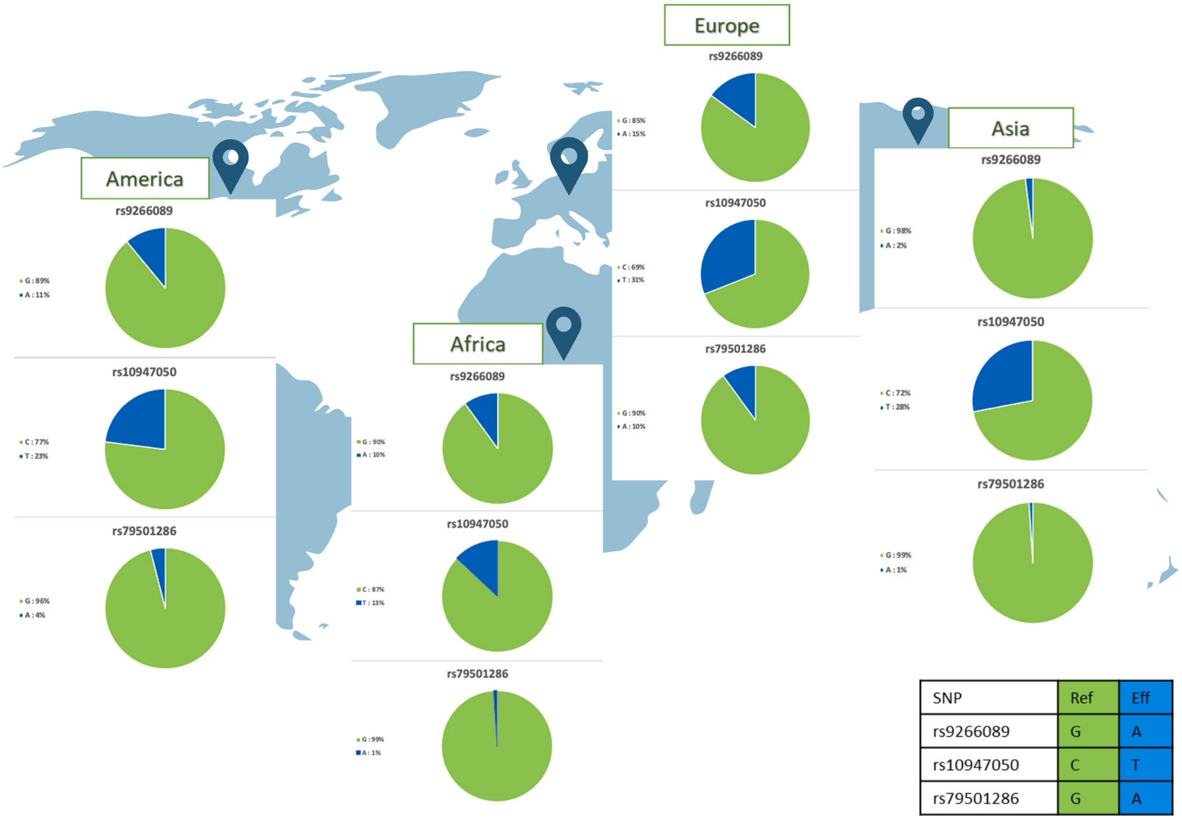
Fig. 3. SNP expression for each genotype across chickenpox disease susceptibility alleles. (**A**) rs9266089 (for *HLA-S* gene); (**B**) rs10947050 (for *HCG4P5* gene); (**C**) rs79501286 (for *ABHD16A* gene).

Table 3

Allele frequencies for each variant associated with chickenpox.

SNP	Position	Gene	Allele		Allele Frequencies			
			Ref	Eff*	AFR	AMR	ASN	EUR
rs9266089	Chr 6 : 31352732	*HLA-S*	G	A	0.10	0.11	0.02	0.15
rs10947050	Chr 6 : 30072514	*HCG4P5*	C	T	0.13	0.23	0.28	0.31
rs79501286	Chr 6 : 31475135	*ABHD16A*	G	A	0.01	0.04	0.01	0.10

AFR, Africa; AMR, America; ASN, Asian; EUR, Europe; Ref, Reference; Eff, Effect allele; AFR, AMR, ASN, EUR, extracted from Haploreg version 4.1 (https://pubs.broadinstitute.org/mammals/haploreg/haploreg.php). *Effect alleles were defined as alleles associated with higher expression of *HLA-S, HCG4P5* and *ABHD16A*.



SNP	Ref	Eff
rs9266089	G	A
rs10947050	C	T
rs79501286	G	A

Fig. 4. Distribution of allele frequencies of the three variants (rs9266089, rs10947050, rs79501286) affecting *HLA-S, HCG4P5*, and *ABHD16A* among different populations across continents. This figure was created with BioRender.com under agreement number "CX24TS7TWB ".

peripheral blood, which plays a central role in the "primary immune response" of cell-mediated immunity. Furthermore, *ABHD16A* is a protein with a hydrolase/domain that has been linked to familial metabolic disorders [16]. We noted the important roles played by the *ABHD16A,* human *ABHD16A* is also known as human leukocyte antigen B (*HLA-B*). Another observation of interest for gene *ABHD16A* supported that the gene could play an important role and function in immune regulation processes [17]. Moreover, *ABHD16A* is associated with immunoregulation [18], kawasaki disease, and coronary artery aneurysms [19]. Another gene, such as *HCG4P5* is also known as *HLA* complex group 4 pseudogene 5 [ENSG00000227766] *HLA* class I and functions for interferon signaling [20,21]. It is well known that Interferon plays a critical role in the human immune response [2]. Human genomic expression levels and patterns in different tissues might be critical for evaluating the susceptibility, symptoms, and outcome of disease, including for chickenpox infectious disease. We sincerely expect the results could be of immense essential in understanding the genetic basis of chickenpox infectious disease. Therefore, we strongly recommend that future research investigate these variants' functions, interactions, and mechanisms in the human body.

Multiple clinical characteristics and susceptibility loci were associated with VZV infection. *HLA* is located on chromosome 6 and is associated with VZV-caused herpes zoster [10]. *ABHD16A* is a member of the alpha-beta hydrolase domain-containing (*ABHD*) family of serine hydrolase metabolism, also referred to as leukocyte antigen-B- (*HLA-B*)-associated transcript 5 (*HLA-B*-associated transcript 5, *BAT5*), which is associated with familial hereditary metabolic disease [16]. We examined chickenpox-associated genetic variants using publicly accessible databases, such as the GWAS catalog [22], Haploreg version 4.1 [23], GTEx portal [24], and Ensembl [25]. Multiple studies have utilized these integration techniques to identify the *TMPRSS2* gene [26], the *ACE2* gene for COVID-19 [26], and a variant for Sjogren syndrome [27]. *HLA-S*, *HCG4P5*, and *ABHD16A* influence the occurrence of chickenpox, which causes fever, headache, backache, chills, nausea, vomiting, and a rash throughout the skin [3,4]. Currently, it is interesting to evaluate and use the variants for a variety of purposes. The variants are not only can guide the susceptibility of the diseases but also can be utilized to drive drug discovery [28–31].

In our study, we highlighted that the variants of rs9266089, rs10947050, and rs79501286 may affect the expression of *HLA-S, HCG4P5,* and *ABHD16A***,** respectively. Based on our study, the AA genotype of rs9266089 had a higher *HLA-S* expression in the skin tissue than the GA (intermediate) and GG (lowest) genotypes (Table 3 and Fig. 3). The A allele frequency of SNP rs9266089 was 2% in Asians compared to Europeans, Americans, and Africans (Fig. 4). At rs10947050, the TT genotype had a higher *HCG4P5* skin expression than the heterozygous CT (intermediate) and homozygous CC (lowest) among European, American, and African populations. African populations had the lowest frequency of the T allele of SNP rs10947050, at 13%, compared to European, American, and Asian populations (Fig. 4). While the AA genotype at rs79501286 in *ABHD16A* had a higher expression in the skin than the heterozygous GA genotype (middle expression) and homozygous GG genotype (lowest expression) (Table 3 and Fig. 3), which were related to chickenpox illness susceptibility. The A allele frequency of SNP rs79501286 showed the lowest yield in Asian and African populations at 1% compared to European and American populations (Fig. 4)~~.~~ Asian individuals had the lowest allele frequencies in the skin for the variations (rs9266089, rs10947050, and rs79501286) related to chickenpox illness compared to European, American, and African groups. Consequently, it can be suggested that European, American, and African populations may have more individuals with the trend of elevated expression of *HLA-S, HCG4P5*, and *ABHD16A* compared to Asians, which may contribute to an increase in the susceptibility to chickenpox disease.

Overall, the allele frequencies of the variants (rs9266089, rs10947050, and rs79501286) that are associated with chickenpox disease showed the lowest allele frequencies in the Asian population. Thus, it can be assumed that European, American, and African populations may have more individuals with increased expression of *HLA-S, HCG4P5*, and *ABHD16A* compared to Asians, which may contribute to an increase in susceptibility to chickenpox disease. This study identified gene variants *in silico*, which must be validated in the clinical implementation of the varicella-zoster virus. However, we believe that meta-analysis based approach is essential to sharpen our prediction as a more comprehensive range of genetic variant data associated with chickenpox becomes available. In addition, further research is needed to validate our findings using functional studies. The validation aims to ascertain whether these variants have a critical role and could harbor a clinically actionable variant for the varicella-zoster virus.

9. Conclusion

Our study emphasizes the investigation of genetic variations influencing chickenpox that revealed the highest expression of *HLA-S*, *HCG4P5*, and *ABHD16A* in the skin tissue. These genes affected the chickenpox risk in individuals and populations. We discovered that the variants rs9266089, rs10947050, and rs79501286 affected *HLA-S*, *HCG4P5*, and *ABHD16A* expression, respectively. Allele frequency is crucial when predicting chickenpox illness susceptibility on multiple continents. Future avenues of research would aim to verify the impact of these variants and their clinical actionability in the diagnosis and management of chickenpox patients.

Author contribution

L.M.Irham and W. Adikusuma conceived and designed the study. L.M.Irham and W. Adikusuma performed the computational analysis. L.M. Irham wrote the manuscript. L.M.Irham provided the funding. A R Afief., A.N Puspitaningrum., Lolita., M.A.Sarasmita and H Dania.,S Khairi, Gina N Djalilah, B Djaka Purwanto and Rocky Cheung revised the manuscript. All authors have read and approved the manuscript and have made significant contributions to this study.

Declaration of competing interest

The authors disclose no conflict.

Data availability

Data will be made available on request.

Acknowledgment

This project was supported by grants from Riset Muhammadiyah (RisetMu) Batch VI 2022 Majelis DIKTI-LITBANG PP Muhammadiyah & Universitas Ahmad Dahlan (No:1687.386/PD/I.3/D/2022).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.bbrep.2022.101419.

References

[1] F. Ayoade, S. Kumar, Varicella zoster, in: *StatPearls*, StatPearls Publishing Copyright © 2022, StatPearls Publishing LLC, Treasure Island (FL), 2022.

[2] N. Kopitar-Jerala, The role of interferons in inflammation and inflammasome activation, Front. Immunol. 8 (2017) 873, https://doi.org/10.3389/fimmu.2017.00873.

[3] Z.S. Moore, J.F. Seward, B.M. Watson, T.J. Maupin, A.O. Jumaan, Chickenpox or smallpox: the use of the febrile prodrome as a distinguishing characteristic, Clin. Infect. Dis. 39 (2004) 1810–1817, https://doi.org/10.1086/426026.

[4] A.A. Gershon, M.D. Gershon, Pathogenesis and current approaches to control of varicella-zoster virus infections, Clin. Microbiol. Rev. 26 (2013) 728–743, https://doi.org/10.1128/cmr.00052-13.
[5] S. Daulagala, F. Noordeen, Epidemiology and factors influencing varicella infections in tropical countries including Sri Lanka, Virusdisease 29 (2018) 277–284, https://doi.org/10.1007/s13337-018-0459-z.
[6] C.K. Fairley, E. Miller, Varicella-zoster virus epidemiology–a changing scene? J. Infect. Dis. 174 (Suppl 3) (1996) S314–S319, https://doi.org/10.1093/infdis/174.supplement_3.s314.
[7] M. Socan, N. Berginc, J. Lajovic, Varicella susceptibility and transmission dynamics in Slovenia, BMC Publ. Health 10 (2010) 360, https://doi.org/10.1186/1471-2458-10-360.
[8] S. Lolekha, W. Tanthiphabha, P. Sornchai, P. Kosuwan, S. Sutra, B. Warachit, S. Chup-Upprakarn, Y. Hutagalung, J. Weil, H.L. Bock, Effect of climatic factors and population density on varicella zoster virus epidemiology within a tropical country, Am. J. Trop. Med. Hyg. 64 (2001) 131–136, https://doi.org/10.4269/ajtmh.2001.64.131.
[9] M.J. Levin, Varicella-zoster virus and virus DNA in the blood and oropharynx of people with latent or active varicella-zoster virus infections, J. Clin. Virol. 61 (2014) 487–495, https://doi.org/10.1016/j.jcv.2014.09.012.
[10] D.R. Crosslin, D.S. Carrell, A. Burt, D.S. Kim, J.G. Underwood, D.S. Hanna, B. A. Comstock, E. Baldwin, M. de Andrade, I.J. Kullo, et al., Genetic variation in the HLA region is associated with susceptibility to herpes zoster, Gene Immun. 16 (2015) 1–7, https://doi.org/10.1038/gene.2014.51.
[11] A.M. Arvin, Varicella-zoster virus, Clin. Microbiol. Rev. 9 (1996) 361–381, https://doi.org/10.1128/cmr.9.3.361.
[12] A.C. Nica, E.T. Dermitzakis, Expression quantitative trait loci: present and future, Philos. Trans. R. Soc. Lond. B Biol. Sci. 368 (2013), 20120362, https://doi.org/10.1098/rstb.2012.0362.
[13] P. Meysman, B. Ogunjimi, S. Naulaerts, P. Beutels, V. Van Tendeloo, K. Laukens, Varicella-zoster virus-derived major histocompatibility complex class I-restricted peptide affinity is a determining factor in the HLA risk profile for the development of postherpetic neuralgia, J. Virol. 89 (2015) 962–969, https://doi.org/10.1128/jvi.02500-14.
[14] S.Y. Choo, The HLA system: genetics, immunology, clinical testing, and clinical implications, Yonsei Med. J. 48 (2007) 11–23, https://doi.org/10.3349/ymj.2007.48.1.11.
[15] H. Fouladseresht, A. Safa, S. Khosropanah, M. Doroudchi, Increased frequency of HLA-A*02 in patients with atherosclerosis is associated with VZV seropositivity, Arch. Physiol. Biochem. 127 (2021) 351–358, https://doi.org/10.1080/13813455.2019.1640253.
[16] A. Yahia, L.E.O. Elsayed, R. Valter, A.A.A. Hamed, I.N. Mohammed, M.A. Elseed, M.A. Salih, T. Esteves, N. Auger, R. Abubaker, et al., Pathogenic variants in ABHD16A cause a novel psychomotor developmental disorder with spastic paraplegia, Front. Neurol. 12 (2021), 720201, https://doi.org/10.3389/fneur.2021.720201.
[17] J. Xu, W. Gu, K. Ji, Z. Xu, H. Zhu, W. Zheng, Sequence analysis and structure prediction of ABHD16A and the roles of the ABHD family members in human disease, Open Biol. 8 (2018), https://doi.org/10.1098/rsob.180017.
[18] S.S. Kamat, K. Camara, W.H. Parsons, D.H. Chen, M.M. Dix, T.D. Bird, A.R. Howell, B.F. Cravatt, Immunomodulatory lysophosphatidylserines are regulated by ABHD16A and ABHD12 interplay, Nat. Chem. Biol. 11 (2015) 164–171, https://doi.org/10.1038/nchembio.1721.
[19] Y.Y. Hsieh, Y.J. Lin, C.C. Chang, D.Y. Chen, C.M. Hsu, Y.K. Wang, K.H. Hsu, F. J. Tsai, Human lymphocyte antigen B-associated transcript 2, 3, and 5 polymorphisms and haplotypes are associated with susceptibility of Kawasaki disease and coronary artery aneurysm, J. Clin. Lab. Anal. 24 (2010) 262–268, https://doi.org/10.1002/jcla.20409.
[20] F. Carlini, V. Ferreira, S. Buhler, A. Tous, J.F. Eliaou, C. René, J. Chiaroni, C. Picard, J. Di Cristofaro, Association of HLA-A and non-classical HLA class I alleles, PLoS One 11 (2016), e0163570, https://doi.org/10.1371/journal.pone.0163570.
[21] C. Lundtoft, P. Pucholt, J. Imgenberg-Kreuz, J. Carlsson-Almlöf, M.L. Eloranta, A. C. Syvänen, G. Nordmark, J.K. Sandling, I. Kockum, T. Olsson, et al., Function of multiple sclerosis-protective HLA class I alleles revealed by genome-wide protein-quantitative trait loci mapping of interferon signalling, PLoS Genet. 16 (2020), e1009199, https://doi.org/10.1371/journal.pgen.1009199.
[22] A. Buniello, J.A.L. MacArthur, M. Cerezo, L.W. Harris, J. Hayhurst, C. Malangone, A. McMahon, J. Morales, E. Mountjoy, E. Sollis, et al., The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019, Nucleic Acids Res. 47 (2019) D1005–d1012, https://doi.org/10.1093/nar/gky1120.
[23] L.D. Ward, M. Kellis, HaploReg v4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease, Nucleic Acids Res. 44 (2016) D877–D881, https://doi.org/10.1093/nar/gkv1340.
[24] J. Lonsdale, J. Thomas, M. Salvatore, R. Phillips, E. Lo, S. Shad, R. Hasz, G. Walters, F. Garcia, N. Young, et al., The genotype-tissue expression (GTEx) project, Nat. Genet. 45 (2013) 580–585, https://doi.org/10.1038/ng.2653.
[25] E. Birney, T.D. Andrews, P. Bevan, M. Caccamo, Y. Chen, L. Clarke, G. Coates, J. Cuff, V. Curwen, T. Cutts, et al., An overview of Ensembl, Genome Res. 14 (2004) 925–928, https://doi.org/10.1101/gr.1860604.
[26] L.M. Irham, W.H. Chou, M.J. Calkins, W. Adikusuma, S.L. Hsieh, W.C. Chang, Genetic variants that influence SARS-CoV-2 receptor TMPRSS2 expression among population cohorts from multiple continents, Biochem. Biophys. Res. Commun. 529 (2020) 263–269, https://doi.org/10.1016/j.bbrc.2020.05.179.
[27] A.N. Puspitaningrum, A.P. D, A. W, G.N. Djalilah, H. D, R. M, F. I.N, A.S. M, M. Rezadhini, et al., Integration of genomic databases and bioinformatic approach to identify genomic variants for sjogren's syndrome on multiple continents, Media Farmasi: J. Ilmu Farm.(J. Pharmaceut. Sci.) 19 (2022) 71–81.
[28] L.M. Irham, W. Adikusuma, D.A. Perwitasari, H. Dania, R. Maliza, I.N. Faridah, I. N. Santri, Y.V.A. Phiri, R. Cheung, The use of genomic variants to drive drug repurposing for chronic hepatitis B, Biochem.Biophys.Rep. 31 (2022), 101307, https://doi.org/10.1016/j.bbrep.2022.101307.
[29] A.R. Afief, L.M. Irham, W. Adikusuma, D.A. Perwitasari, A. Brahmadhi, R. Cheung, Integration of genomic variants and bioinformatic-based approach to drive drug repurposing for multiple sclerosis, Biochem.Biophys.Rep. 32 (2022), 101337, https://doi.org/10.1016/j.bbrep.2022.101337.
[30] L.M. Irham, H.S.-C. Wong, W.-H. Chou, W. Adikusuma, E. Mugiyanto, W.-C. Huang, W.-C. Chang, Integration of genetic variants and gene network for drug repurposing in colorectal cancer, Pharmacol. Res. 161 (2020), 105203, https://doi.org/10.1016/j.phrs.2020.105203.
[31] L.M. Irham, W. Adikusuma, D.A. Perwitasari, Genomic variants-driven drug repurposing for tuberculosis by utilizing the established bioinformatic-based approach, Biochem.Biophys.Rep. 32 (2022), 101334, https://doi.org/10.1016/j.bbrep.2022.101334.

View publication stats

Ormel et al. *Translational Psychiatry* (2019)9:114
https://doi.org/10.1038/s41398-019-0450-5

Translational Psychiatry

ARTICLE Open Access

The genetics of depression: successful genome-wide association studies introduce new challenges

Johan Ormel[1], Catharina A. Hartman[1] and Harold Snieder[1]

Abstract

The recent successful genome-wide association studies (GWASs) for depression have yielded more than 80 replicated loci and brought back the excitement that had evaporated during the years of negative GWAS findings. The identified loci provide anchors to explore their relevance for depression, but this comes with new challenges. Using the watershed model of genotype–phenotype relationships as a conceptual aid and recent genetic findings on other complex phenotypes, we discuss why it took so long and identify seven future challenges. The biggest challenge involves the identification of causal mechanisms since GWAS associations merely flag genomic regions without a direct link to underlying biological function. Furthermore, the genetic association with the index phenotype may also be part of a more extensive causal pathway (e.g., from variant to comorbid condition) or be due to indirect influences via intermediate traits located in the causal pathways to the final outcome. This challenge is highly relevant for depression because even its narrow definition of major depressive disorder captures a heterogeneous set of phenotypes which are often measured by even more broadly defined operational definitions consisting of a few questions (minimal phenotyping). Here, Mendelian randomization and future discovery of additional genetic variants for depression and related phenotypes will be of great help. In addition, reduction of phenotypic heterogeneity may also be worthwhile. Other challenges include detecting rare variants, determining the genetic architecture of depression, closing the "heritability gap", and realizing the potential for personalized treatment. Along the way, we identify pertinent open questions that, when addressed, will advance the field.

Introduction

Major depressive disorder (MDD, henceforth: depression) is a common mental disorder. In most Western countries, MDD has a 1-year prevalence of ~5%[1,2] and a lifetime prevalence of ~15%[3]. Depression is in the top three of the leading causes of years lived with disability and a significant contributor to premature mortality due to suicide[4,5]. Twin-based heritability estimates typically fluctuate around 35%. Candidate gene studies have not implicated replicable gene variants[6] and—until very recently—genome-wide association studies (GWASs) had neither[7,8].

This long drought without much progress in understanding the genetics of depression has recently ended. The CONVERGE consortium focused on a homogeneous subtype of carefully phenotyped recurrent, severe depression, and identified two genetic loci exceeding genome-wide significance levels[9]. Shortly thereafter, GWASs that used the alternative strategy of increasing sample size while using a more lenient, easy to measure, depression phenotype (i.e., "minimal" phenotyping) have identified additional genetic variants[10–12], culminating in the most recent successes of GWASs conducted in the UK Biobank ($n = 322{,}580$; 16 independent loci associated with broad depression phenotypes)[13], the PGC MDD working group (130,664 MDD cases and 330,470 controls; 44 independent loci)[14], and the Howard et al. meta-meta-

---

Correspondence: Johan Ormel (j.ormel@umcg.nl)
[1]Departments of Epidemiology and Psychiatry, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands

© The Author(s) 2019
Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

analysis of data on 807,553 individuals (246,363 cases and 561,190 controls) from the three largest GWASs of depression using minimal phenotyping[15]. Howard and colleagues identified 102 independent variants of which 87 replicated in an independent sample. Finally, GWAS findings on multiple psychiatric phenotypes were leveraged to identify eight novel independently replicated depression loci[16].

With this successful harvest of GWAS loci for depression, the next wave of challenges has come to the fore. Using the watershed model of genotype–phenotype relationships as a conceptual aid and recent genetic insights on other complex traits such as height, we discuss first why it took so long, followed by seven challenges that need further work to advance the field. The biggest challenge is probably the identification of genetic variants that are causally involved in MDD. These challenges are not unique to depression but apply in varying degrees to highly polygenic conditions in general. Nonetheless, we focus this paper on depression for two reasons: (1) Depression is an interesting mental disorder because of its high prevalence and relatively small twin-based heritability, and (2) It affords a clear focus and obviates the need to include information about many other polygenic conditions. Depression contrasts strongly with human height, a complex trait like depression, which is—unlike depression—highly heritable and measured with less error. Below, we often use height as a contrasting complex trait. Gene finding studies for height have shown much more rapid returns than for depression and may give insight into what we might expect for depression in the near future with much larger sample sizes compensating for its lower heritability and larger measurement error.

Why did it take so long?

Keller and Cannon's watershed analogy of the genotype–phenotype relationship is helpful in explaining why the identification of genetic loci for depression has been extremely difficult (Fig. 1)[17]. Much like the numerous streams of the Amazonian watershed that merge and eventually empty into the Atlantic Ocean, sets of genetic variants influence upstream micro-biological processes (narrowly defined mechanisms; e.g., dopamine transmission) that influence intermediate "downstream" meso-biological processes (e.g., working memory, facial emotion recognition) that in turn affect macro-biological processes (e.g., stress sensitivity, affect regulation), which contribute to the overall fitness of the individual. Gene variants that are directly involved in an upstream process not only affect that upstream process, but indirectly also many processes and traits downstream. If enough noise is present in particular upstream processes, specific behavioral syndromes may arise, such as symptoms of mental disorder. The more upstream (specific, narrow) the (intermediate) phenotype is, the closer its relationship to the genetic variants that affect it, thereby increasing the probability that the associated genetic variants are truly involved in the production of the phenotype.

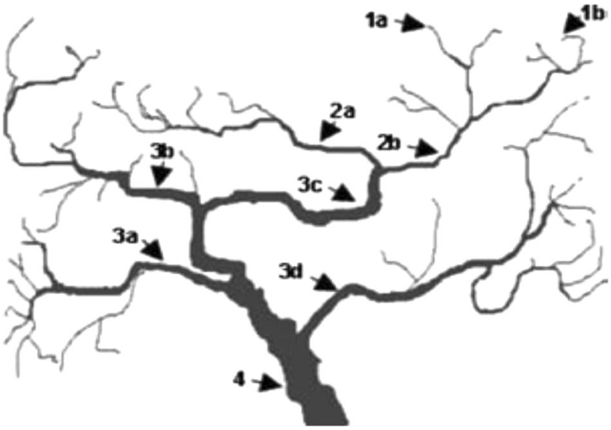


Fig. 1 The watershed model of the pathways connecting upstream genetic variants to downstream depression. Mutations at specific loci (1**a**, 1**b**) disrupt narrowly defined mechanisms such as transmission of dopamine in the prefrontal cortex (2**b**). This and other narrowly defined mechanisms contribute noise to more broadly defined mechanisms, such as working memory, facial emotion recognition (3**c**). The latter mechanisms in conjunction with several other mechanisms (3**a**, 3**b**, 3**d**) affect observable phenotypes, such as stress sensitivity, affect regulation. All tributaries eventually flow into fitness (4). Adapted from Keller & Miller, 2006[17]

Depression is a rather downstream phenotype, almost at the level of fitness (Fig. 1). This feature may explain why it took so long for GWASs of depression to become successful. First, as illustrated by the watershed model of genotype–phenotype relationships, multiple tributaries may lead to the same phenotypic outcome (so-called equifinality), with the implication that depression is etiologically heterogeneous, consistent with current insights. This hampers attempts to identify the genetic loci for depression since all individuals with depression are lumped together, which dilutes the explained variance of a particular variant[18].

Second, depression is not only etiologically but also phenotypically heterogeneous. Approximately 1500 DSM-IV symptom combinations can fulfill the diagnostic criteria[19]. Two patients diagnosed with depression may have very few symptoms in common. In addition, phenotypic heterogeneity will be enhanced by co-occurring psychiatric problems. Two patients having similar depression symptoms may have different comorbid conditions (e.g., anxiety disorder versus substance use disorder). Again, this will dilute the effects of a particular variant.

Third, the upstream processes involved in depression are hardly known, with problematic consequences for the use of "endophenotypes". To date, the endophenotype approach in depression has not been able to live up to its promises. Many depression-related endophenotypes proposed in the past were based on stress research and neuroimaging, and thus still at relatively low levels in the watershed model. Such endophenotypes tend to be as genetically complex as their outcome phenotypes[20].

Fourth, phenotypes downstream in the watershed are more likely to be "constructs" rather than biological "entities", which introduces measurement error. Depression, whether measured by inventories, (semi-)structured interviews, or clinical assessments, is subjective and based on reported and observed affect, cognition, and behavior. Human height, by comparison, is a much simpler phenotype that can be obtained objectively and reliably. Moreover, measurement of depression is complicated by its mixed course, ranging from a single lifetime episode to chronic-recurrent. Even during an episode symptoms often fluctuate. This complicates measurement since retrieving "the most severe episode" from memory implies retrospective assessment with all its reliability problems. For example, it has been shown that SNP heritability is considerably higher for emotional problems when focusing on the stable variance over time compared to single measures per time point, which further illustrates the relative advantage of height, which remains constant over the adult life course, over depression as a phenotype for genetic studies.

Challenges pertinent to the interpretation and utility of detected genetic variants

With the recent GWAS successes of identifying more than 80 replicated loci, the next wave of challenges for the genetic study of depression come to the fore.

Challenge 1. Prioritizing likely causal genes for functional follow up

An important characteristic of GWASs is that the identified variants merely flag genomic regions without necessarily providing a direct link to the underlying biological mechanisms[21]. In addition, the identified variant may not be (directly) causal to the phenotype of interest but to other phenotypes that are strongly associated with the phenotype of interest, including comorbid conditions and intermediate traits in the causal pathway leading to the final outcome. Furthermore, effect sizes of individual genetic variants are typically very small (although this does not rule out that effect sizes on currently unknown micro-biological phenotypes higher in the watershed can be large)[22]. All in all, the selection of the most promising signals and the discovery of their functional consequences represent a major challenge. Given the costs and difficulties of conducting functional studies, prioritization of likely causal genes is very important. For the approximately 80+ depression loci this is a formidable task. To date, bioinformatics analyses have been the main strategy.

Fine-mapping of identified loci is typically used as a first step to limit the "credible set of SNPs" that likely include the causal variant(s) responsible for the observed GWAS signals. Transethnic differences in linkage disequilibrium can be used to improve its resolution[23]. However, with the exception of the CONVERGE study of recurrently depressed Han Chinese women, depression GWASs have been limited to individuals of European ancestry. Therefore, more GWASs in other ethnic groups are needed to aid fine mapping efforts. Further bioinformatic post-GWAS follow up analyses leverage the fact that there are only two biological mechanisms that can explain true SNP–phenotype associations: (1) the SNP may alter the amino acid coding (i.e., a nonsynonymous SNP) changing the protein structure and potentially its function, alternatively; (2) it may exert its phenotypic effect through influencing the expression of the gene. Therefore, bioinformatic post-GWAS pipelines will check whether GWAS signals will be in high linkage disequilibrium with nonsynonymous SNPs within nearby genes and use publicly available expression Quantitative Trait Loci (eQTL) resources from relevant tissues such as the brain (GTEx, Braineac) or whole blood to check which SNPs in the identified loci are also associated with gene expression[24].

Recent tools integrate evidence from GWAS with eQTL data within a Mendelian randomization framework allowing assignment of likely causality of genes within loci[25–27]. The Mendelian randomization approach is based on the fact that the DNA sequence is fixed, which implies that causation can only flow in one direction, allowing the use of genetic markers as instrumental variables. With the advent of GWAS and the recent explosion in variant discovery, this simple idea has been applied to great effect[28]. It has been used in particular to examine the causality of correlated phenotypes and the origins of genetic correlations (see also Challenge 4). But it can also be used to explore the causal role of genes within identified genetic loci for the phenotype of interest[25].

Mendelian randomization in its most basic form is summarized in Fig. 2 where the causal relationship between an exposure (e.g., obesity) and an outcome (e.g., depression) is investigated (association 1 in Fig. 2) using genetic variants known for influencing the exposure (obesity) (association 2) as an instrumental variable, by estimating the association between the genetic variants and the outcome (depression) (association 3)[27].

Mendelian randomization has three important assumptions (see Fig. 3). First, the genetic variants should have a robust and strong relationship with the exposure (a). Second, the variants must not be associated with factors that confound the association between exposure and outcome (b). Third, the genetic markers for the exposure only influence the outcome through their effect on the exposure and not through any other pathway (c). Although the latter two assumptions are hard to prove, many Mendelian randomization sensitivity analyses have recently been developed that are less reliant on these assumptions[29].

Further bioinformatic analyses may explore whether genes within the GWAS loci are preferentially expressed in certain tissues or enriched in certain networks and pathways, and whether these genes are targets of existing (e.g., psychiatric) medications[22,30]. Ultimately, unequivocal evidence of underlying mechanisms will have to come from functional studies such as the one showing a role in synapse pruning for the *C4A* gene in schizophrenia[31]. Developing and increasing throughput of either cell-based or animal model assays to investigate the many GWAS loci for function will be one of the main challenges for the immediate future.

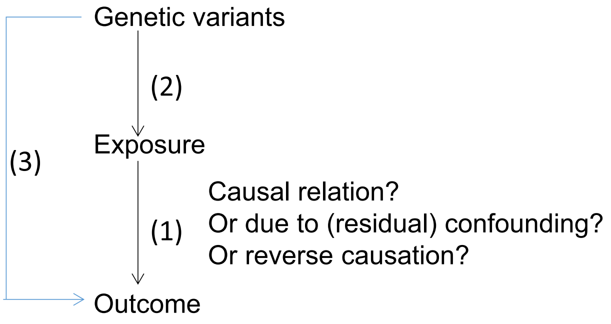


Fig. 2 Framework of a Mendelian randomization study in its most basic form. Adapted from Verduijn, 2010[27]

Challenge 2. Finding rare and more common variants

Given the relationship between sample size and the number of detected loci, it is to be expected that larger sample sizes will identify additional loci. These will include common variants with (even) small(er) effects and probably rare variants of moderate-to-large effect although their role in depression is currently unknown[32]. The brief history of GWASs of, for example, schizophrenia, supports the expectation that with larger sample sizes rare variants can be discovered. In general, the contribution of rare variations of strong effect tends to be larger for early-onset, highly heritable, severe (e.g., neurodevelopmental disorders, including schizophrenia) disorders and lesser for disorders that are less heritable, less severe and with a later onset, such as depression[30]. But that does not exclude a role for rare variants in depression.

The support that rare variants may play a role in depression also comes from large GWASs of complex traits outside psychiatry, most notably human height[33,34]. Some rare variants (minor allele frequencies [MAFs] 0.8–2.1%) had large effects, implicating a 2 cm difference in height. The explained variance of genetic variants is a simple function of both effect size and MAF. As such, despite their much larger effect size, the rare height-associated genetic variants each explain, on average, similar amounts of variation at the population level as common variants. The much lower effect size of common variants is “compensated” by their much higher frequency.

Fortunately, costly whole genome or exome sequencing may not be necessary to find rare variants, as large GWAS sample sizes of a million individuals or more, imputed to very large sequenced reference samples will offer sufficient resolution and power in the low frequency range and

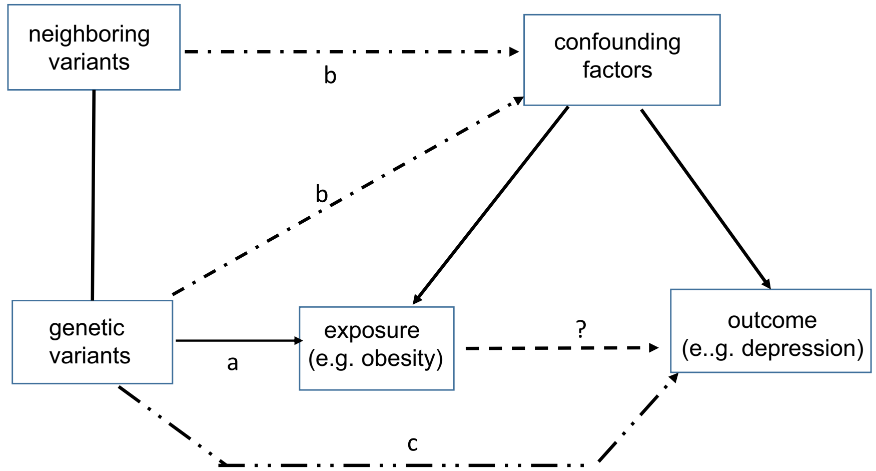


Fig. 3 Assumptions made in Mendelian randomization: (**a**) presence of a robust association between genetic variants and exposure (here depression), (**b**) absence of (direct/indirect) association between generic variant and confounding factors, and (**c**) absence of other pathways between genetic variants and outcome. Adapted from Verduijn, 2010[27]

are now increasingly feasible[35,36]. Furthermore, the strong phenotypic and genetic correlation between depression and other mental disorders may be leveraged to improve power and identify both additional common and novel rare variants for depression[16].

Challenge 3. Establishing the genetic architecture of depression

Initially, GWASs of complex traits operated from the simple common disease–common variant (CDCV) model, positing that a moderate number (<100) of gene variants of intermediate frequency (MAF > 5%) with small-to-moderate effect (OR > 1.5) account for the heritability of the trait. The results of recent GWAS studies of complex traits, including height, have shown this CDCV model to be wrong for most types of complex traits[37–39]. Four alternative models of genetic architecture have been proposed: (i) the *infinitesimal* model assuming that heritability is due to a large number (»100) of small-effect common variants, (ii) the *rare allele* model, assuming that heritability is due to a large number of rare variants with relatively large-effects (including copy number variants), (iii) the *broad-sense heritability* model, assuming that in addition to additive effects of common variants, heritability is due to rare variants, non-additive GxG (dominance, epistasis) and GxE interactions as well as epigenetic effects, and (iv) the *omnigenic* model hypothesizing that the genetic architecture of complex traits is characterized by very large numbers of peripheral, more general genes and a limited number of "core" genes, assumed to be more disease specific[37,40].

Apart from the yet to be detected occasional rare variants, the infinitesimal model may provide a good approximation for a highly polygenic disorder like depression. Regarding non-additive GxG and GxE interactions, decisive evidence is lacking as to date GWASs have not been designed and powered to detect GxG and GxE interactions, neither for depression, nor for other mental disorders or height. Recent studies were unable to show a sizable influence of dominance for a wide range of complex traits[41]. Previous attempts to study GxE on the basis of single candidate genes[42,43] are now being replaced by the use of polygenic risk scores (e.g., refs. [44,45]). which may offer more promise for detection of GxE interaction, although the polygenic risk scores will only provide a general measure of genetic susceptibility for depression that does not directly indicate underlying mechanisms. Moreover, the use of polygenic risk scores neglect the possibility that the impact of GxE interaction is not consistent across different genes or pathways; some genetic pathways may show GxE interaction but others may not or in the reverse direction. Nonetheless, it is likely that these interaction studies will eventually be successful given, for example, "indirect" evidence of substantial GxE interactions between personality traits and environmental exposures[46,47]. Regarding the omnigenic model, Wray and colleagues have recently argued that, although intuitively appealing, there is insufficient empirical evidence for the existence of its hypothesized core genes[18].

Irrespective of the correct model for the genetic architecture, it must reserve an honorable seat for environmental influences. In addition to the aforementioned interactions, two epidemiological observations support a substantial environmental component in the etiology of depression: Although depression can be found everywhere, substantial national and regional variations in prevalence exist[3,48]. Because of the ultimately subjective nature of the measurement process, these prevalence differences are hard to interpret. Furthermore, strong effects on risk for depression have been documented for long-term difficulties (e.g., taking care of a dementing partner, persistent unemployment, victim of chronic bullying) which in addition moderate the depression risk of stressful life events (e.g., acute illness of child, let down by friend)[46,49–51]. The latter short-term factors are typically thought to play a role in the timing of depression onsets pushing susceptible individuals across the diagnostic threshold. Although such environmental effects partly reflect depression's genetic background (i.e., gene–environment correlation); they are nonetheless likely to have an extra, additive, contribution to, as well as interaction with, genetic risk in explaining the depression phenotype.

Challenge 4. Genetic pleiotropy and unraveling causal relationships with other traits

Pleiotropy is the phenomenon whereby a genetic variant influences two or more phenotypes[52]. In line with high comorbidity across mental disorders, GWAS findings also indicate substantial genetic overlap, although the extent thereof was perhaps unexpected[53]. SNP-based genetic correlations (r_g) between depression and other mental disorders are substantial[14]. Pleiotropy is not specific to psychiatry and has also been shown between depression and somatic conditions[14,54,55].

Part of the genetic correlations may derive from symptom/diagnostic overlap, comorbidity or can even be an artefact of diagnostic misclassification. It is useful to make a distinction between heterogeneity due to different symptom patterns that all meet MDD's diagnostic criteria and heterogeneity due to comorbidity with other mental disorders, a common phenomenon. Genetic correlations may also reflect a common cause. For example, virtually all mental disorders involve sensitivity to stressful situations which is why individual differences in appraisal of and coping with stressful experiences have an impact on the severity of their manifestation. This commonality implies that genetic variants that influence appraisal and coping may turn up in GWASs of these mental disorders, although they are at best causally involved in a generic way, not strictly part of the specific disorder's pathophysiology. In the watershed model, these shared appraisal-coping or executive functioning related variants are probably located relatively upstream, with many confluences downstream. The higher hierarchical position makes it "easier" to be involved in multiple phenotypes than variants more downstream.

A unique feature of genetic as opposed to classic epidemiological associations that aids in unraveling causality is that confounders (or third variables) that influence both genetic markers and outcome phenotypes do not exist (DNA is fixed). Mendelian randomization can thus also be used to disentangle the causes of correlated traits using genetic markers to distinguish between alternative causal explanations such as reverse causation and shared causes[56]. One important requirement for effective Mendelian randomization is the availability of a sufficient number of genetic markers associated with the exposure as their combined effect determine the strength of the instrumental variable. This means that distinguishing causality between depression and associated traits using Mendelian randomization has only recently become possible. The continued discovery of additional genetic variants for depression will be important to improve their strength as an instrumental variable and hence the power of Mendelian randomization analyses in order to distinguish between alternative causal pathways, thereby providing fresh clues to old questions.

Challenge 5. Closing the "heritability gap"

Genetic variants detected by GWAS typically explain only a fraction of the total family- or twin-based heritability. This has become known as the missing heritability problem or heritability gap[41,57]. Recently introduced methodology now also allows calculation of the SNP- or chip-based heritability (h^2_{SNP}), which is the proportion of phenotypic variance jointly accounted for by all variants on a standard GWAS chip. h^2_{SNP} provides an upper bound estimate of the genetic effects that could be detected by a (well-powered) GWAS[58]. The remainder is likely due to rarer and structural variants that have until recently not been captured by regular GWAS arrays[35]. The latest PGC depression GWAS estimates this h^2_{SNP} at ~9%, which is only around a quarter of the heritability based on twin and family studies of ~35%. However, measurement error and heterogeneity in phenotype definition between the different PGC cohorts may explain part of the difference in heritability as the CONVERGE study found a h^2_{SNP} between 20 and 29% within their cohort of carefully assessed women with early-onset recurrent depression.

Another part of the heritability gap may be attributed to potentially inflated twin heritability estimates caused by gene by common (C) (or shared) environment (GxC) interactions[45]. That is, genetic effects depend on environmental factors shared by twins that grow up in the same family but not by unrelated individuals in the GWAS samples. The statistical models used in twin studies to estimate twin-based heritability fully attribute the joint effect (GxC) to the genetic component, thus inflating heritability estimates and reducing the contribution of shared environment. The GxC explanation "solves" not only (part of) the heritability gap but the shared environment paradox as well. This paradox refers to the apparent inconsistency that, in contrast to epidemiological studies, the statistical models used in twin studies typically find hardly any shared-environmental variance while many (distal) environmental risk factors are shared by twins in the same family, e.g., poverty, family instability, child neglect, neighborhood stressors, minority status, SES[59]. As pointed out by Uher and Zwicker[45], this paradox disappears if we realize that the impact of these shared-environmental factors depends on characteristics shared to a larger extent by monozygotic than dizygotic twins, i.e., genetic variants.

The difference between heritability explained by GWAS identified variants and the h^2_{SNP} has also been more aptly termed "hidden" heritability as future larger GWASs are expected to detect additional signals still hidden in the noise. Furthermore, new generations of denser and better-imputed GWAS arrays are expected to capture more rare and structural variants, which will increase the h^2_{SNP} and decrease its gap with the total (potentially inflated) (family and twin-based) heritability[35,57].

Challenge 6. Reducing the phenotypic heterogeneity

A major challenge is the identification of genetic variants that are causally involved in MDD and this is aggravated by the use of minimal phenotyping (based on a few symptoms), which, as recently shown, may have yielded cases unrepresentative of MDD that are enriched by people "with non-specific sub-clinical depressive symptoms and depression secondary to a comorbid disease"[60].

Traditionally, endophenotypes of depression have been used in the hope that this would shorten the distance between genes and phenotype, and consequently reduce genetic heterogeneity. To date, the expectations have not come true as the used endophenotypes were genetically not less complex. However, the recently identified genetic signals may provide new insights into underlying pathophysiological pathways and networks providing clues on more suitable less complex endophenotypes (micro-biological phenotypes) that are at higher levels in the watershed model and thus closer to the genes. Constructs assessed through task performance such as reward sensitivity and attentional biases may have potential utility as endophenotypes if they are associated with genetic signals. Given their transdiagnostic relevance, they may also account for other disorders in which these mechanisms are active as well.

Some argue that another way forward lies in the reduction of phenotypic heterogeneity, by targeting a specific subtype such as early-onset recurrent melancholia or a specific symptom cluster[9]. Others believe that a broader phenotype approach will provide a more tractable target for genetic studies, as this identifies more signals[11]. Findings of the CONVERGE consortium based on a sample size that was much smaller than the other recent depression GWASs suggests that in-depth phenotyping may pay off[9]. In addition to high-quality measurement of depression, this study focused on recurrent severe depression which may be genetically more homogeneous. However, CONVERGE differed in additional aspects from the other MDD GWASs, including the Chinese sample, the focus on women, and analytical approach. We cannot rule out that these differences contributed to CONVERGE's success.

The recent GWASs successes with broad phenotypes of depression and neuroticism[61] seem to support the broad trawl approach[13]. But note that phenotype broadness is subject to more noise because the likelihood that identified loci are not involved in the physiology of the phenotype of interest increases with broadness (see Challenge 4).

In contrast to broader milder depressive states, other clinically recognizable "subtypes" of depression including early-onset recurrent depression and the more severe subtypes of melancholic, bipolar, and psychotic depression may also be less heterogeneous[62–64]. It is also important to keep in mind that even highly homogeneous subtypes defined by behavioral symptoms are still placed low in the watershed model and remain multifactorial, with multiple underlying etiological pathways although they may be less genetically heterogeneous than all "depressions" lumped together.

Novel approaches improving the phenotype definition of depression may be needed. We outline two complementary approaches below that make use of depression measured at the most narrow level of individual symptoms.

Bottom-up: individual symptoms of depression as a starting point

Symptom-specific GWASs to examine their genetic background may be an interesting next step. Previous depression GWASs have used composite scores or diagnostic case-control designs. Based on data from the UK Biobank, Nagel and colleagues[65] showed that the composite score of neuroticism, an important personality trait that partially overlaps with depression[66], directs the focus to genetic variants that affect the majority of aggregated items, i.e., "global variants". The genetic signal of "local" variants, affecting only one or a few of the aggregated items, was severely diluted[65]. Given its multidimensional and heterogeneous nature, it is plausible that symptom-specific GWASs of depression will yield similar findings. Some initial evidence comes from a relatively small sample showing that h^2_{SNP} of four depression symptom components (appetite, depressed mood and anhedonia, insomnia, and anxiety) was different, suggesting the possible merit of more narrowly defined phenotypes[67].

Top-down: use hierarchical dimensional models

Hierarchical dimensional models such as the Hierarchical Taxonomy of Psychopathology (HiTOP)[68] show how psychopathology dimensions can be arranged in a hierarchy, ranging from very broad "spectrum level" dimensions (e.g., distress, thought disorder, disinhibited externalizing, etc.), to more specific clusters of symptoms. For instance, the distress spectrum, one of the five internalizing spectra consists of the lower clusters (subdimensions) of irritability, anhedonia, numbing, physical panic, suicidality, dysphoria, retardation, lassitude, appetite loss, insomnia, and generalized anxiety[69]. These models hypothesize a hierarchy linking spectrum level dimensions with highly pleiotropic variants which subsume lower order symptom clusters with less pleiotropic variant clusters. This phenotypic ordering from narrow to broad may help in the interpretation of genetic findings.

Challenge 7. Personalized treatment

An important issue is the utility of identified variants for individualized treatment of depression. Three applications may be genomic (polygenic) risk prediction, genome editing, and the identification of novel "druggable" targets, respectively. GWAS results can be used for genomic risk prediction. Increased risk could prompt more intensive surveillance or even prophylactic treatments unrelated to a specific causal mechanism (cf. preventive mastectomy in case of high genetic breast cancer risk). The widespread pleiotropy among mental disorder phenotypes (genetic correlations) can be used to improve genomic risk prediction, so that this might benefit personalized medicine[70]. However, genetic risk-driven prophylactic treatment is not realistic yet, given the low effect size of SNP-based genetic predictors[70].

Genome-editing technology such as CRISPR/Cas9 may make it possible to change or disable genes in living cells in a precise, cheap, and fast way by cutting, replacing, or adding pieces from the DNA[71,72]. However, it is highly doubtful whether this will ever have relevance for "fixing" depression-associated genetic variants due to their small effect sizes, unknown individual relevance (from population to individual), causal relevance (too upstream), and unwanted "side" effects of genome-editing (genetic pleiotropy).

With regard to pharmacotherapy, virtually all currently used drugs in psychiatry have their origins in chance findings in the previous century, while rational approaches to develop new pharmacological treatments have mostly not paid off. The recent findings from the PGC depression GWAS indicated that lead SNPs in some loci were within genes known to play a role in neuronal development, synaptic function, transmembrane adhesion complexes, and/or regulation of gene expression in brain. In addition, genes that are targets of antidepressant medications were strongly enriched for depression-associated signals, which may indicate pharmacotherapeutic relevance[73]. In addition, some identified loci were associated with clinical features of depression including early-onset, recurrence, and severity, and implicated prefrontal and anterior cingulate cortex in the pathophysiology of depression (brain regions showing MRI anatomical and functional differences between MDD cases and controls). Thus, current genomic findings may have substantial potential for the development of new depression medications. The future will tell.

However, the complex genetic nature of depression raises the question for whom drugs developed on the basis of GWAS findings will work. For a highly complex trait-like depression, each individual probably carries a unique combination of protective and risk alleles (see ref. [18] for an illustration). The more polygenic a trait, the more combinations of these sets of alleles are possible, implying that each individual is likely to have a different combination, including affected individuals whose symptom levels have crossed the diagnostic threshold. This explains why most mental disorders are highly heritable but only weakly inheritable[22]. Due to genetic recombination, the probability that a child will inherit a mix of alleles from an affected parent resulting in a genetic risk sufficiently high to also pass the diagnostic threshold remains fairly small. Thus, effect sizes derived from GWASs of a genetic variant with depression should be interpreted in the context of an averaged background; in individual carriers the contribution of a certain variant may be much larger. A particular drug can be effective only in the subgroup of individuals that share the genetic variant and pathway targeted by the drug. The effectiveness of a drug in the individual case thus depends on the number of possible combinations in which the particular variant is a necessary component to become depressogenic. Ultimately, precision medicine for highly polygenic disorders like depression may depend on successfully matching these unique individual genomic profiles to drug treatments[18].

Conclusion

Clearly, the excitement for investigating the genetic background of depression has returned. Prospects to unravel the pathogenesis and etiology of depression and rationally develop pharmacotherapies are better than ever. Now that depression-associated genetic variants have been found, the next wave of challenges has taken center stage. It is likely that larger samples will identify additional common variants and the first batch of rare variants, which will reveal the genetic architecture of depression, its subtypes and broader phenotypes. It is also likely that taking common (C) (or shared) environment (GxC) interactions into account will contribute to unearthing part of the missing heritability[45]. The GxC explanation not only bridges (part of) the heritability gap but also may explain the shared environment paradox.

The largest challenges will be to identify the causal variants of depression itself and to determine which variants merely correlate with the depression phenotype because they are causally involved in its determinants. In addition to bioinformatics, Mendelian randomization may help to resolve this causal web and will become increasingly effective with the identification of additional genetic variants of depression and related phenotypes. To date, Mendelian randomization has almost exclusively been used to aid in the interpretation of genetic correlations between correlated phenotypes, but it may also assist in distinguishing causal from pleiotropic depression variants. Another approach to clarify the causal status of genetic variants might be comparing markers associated with narrowly (e.g., early-onset recurrent melancholic depression) versus broadly defined depression phenotypes. Likewise, comparing the variants associated with different levels of the "depression hierarchy" (symptoms, clusters of symptoms, internalizing dimensions, general psychopathology factor) might be informative. Finally, the fact that highly polygenic traits are only weakly inheritable despite substantial heritability[22] could explain the seeming paradox that antidepressants may benefit individual patients enormously despite their modest average efficacy on a group level compared to pill-placebo[74].

Conflict of interest

The authors declare that they have no conflict of interest.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 29 January 2019 Accepted: 13 February 2019
Published online: 15 March 2019

References

1. Steel, Z. et al. The global prevalence of common mental disorders: a systematic review and meta-analysis 1980–2013. *Int. J. Epidemiol.* **43**, 476–493 (2014).
2. Ferrari, A. J. et al. Burden of depressive disorders by country, sex, age, and year: findings from the global burden of disease study 2010. *PLoS Med.* **10**, e1001547 (2013).
3. Bromet, E. et al. Cross-national epidemiology of DSM-IV major depressive episode. *BMC Med.* **9**, 90 (2011).
4. Erskine, H. E. et al. A heavy burden on young minds: the global burden of mental and substance use disorders in children and youth. *Psychol. Med.* **45**, 1551–1563 (2015).
5. Whiteford, H. A., Ferrari, A. J., Degenhardt, L., Feigin, V. & Vos, T. The global burden of mental, neurological and substance use disorders: an analysis from the global burden of disease study 2010. *PLoS ONE* **10**, e0116820 (2015).
6. Bosker, F. J. et al. Poor replication of candidate genes for major depressive disorder using genome-wide assoiation data. *Mol. Psychiatry* **16**, 516–532 (2011).
7. Wray, N. R. et al. Genome-wide association study of major depressive disorder: new results, meta-analysis, and lessons learned. *Mol. Psychiatry* **17**, 36–48 (2012).
8. Sullivan, P. & 96 Psychiatric Genetics Investigators. Don't give up on GWAS. *Mol. Psychiatry* **17**, 2–3 (2012).
9. Cai, N. et al. Sparse whole-genome sequencing identifies two loci for major depressive disorder. *Nature* **523**, 588–591 (2015).
10. Okbay, A. et al. Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. *Nat. Genet.* **48**, 624–633 (2016).
11. Hyde, C. L. et al. Identification of 15 genetic loci associated with risk of major depression in individuals of european descent. *Nat. Genet.* **48**, 1031–1036 (2016).
12. Direk, N. et al. An analysis of two genome-wide association meta-analyses identifies a new locus for broad depression phenotype. *Biol. Psychiatry* **82**, 322–329 (2017).
13. Howard, D. M. et al. Genome-wide association study of depression phenotypes in UK Biobank identifies variants in excitatory synaptic pathways. *Nat. Commun.* **9**, 1470 (2018).
14. Wray, N. R. et al. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat. Genet.* **50**, 668–681 (2018).
15. Howard, D. M. Genome-wide meta-analysis of depression in 807,553 individuals identifies 102 independent variants with replication in a further 1,507,153 individuals. *Nat. Neurosci.* **22**, 343–352 (2019).
16. Amare, A. T. et al. Bivariate genome-wide association analyses of the broad depression phenotype combined with major depressive disorder, bipolar disorder or schizophrenia reveal eight novel genetic loci for depression. *Mol. Psychiatry* https://doi.org/10.1038/s41380-018-0336-6 (2019).
17. Keller, M. C. & Miller, G. Resolving the paradox of common, harmful, heritable mental disorders: which evolutionary genetic models work best? *Behav. Brain Sci.* **29**, 385–404 (2006).
18. Wray, N. R., Wijmenga, C., Sullivan, P. F., Yang, J. & Visscher, P. M. Common disease is more complex than implied by the core gene omnigenic model. *Cell* **173**, 1573–1580 (2018).
19. Ostergaard, S. D., Jensen, S. O. W. & Bech, P. The heterogeneity of the depressive syndrome: when numbers get serious. *Acta Psychiatr. Scand.* **124**, 495–496 (2011).
20. Franke, B. et al. Genetic influences on schizophrenia and subcortical brain volumes: large-scale proof of concept. *Nat. Neurosci.* **19**, 420–431 (2016).
21. Vaez, A. et al. In silico post genome-wide association studies analysis of C-reactive protein loci suggests an important role for interferons. *Circ. Cardiovasc. Genet.* **8**, 487–497 (2015).
22. Visscher, P. M. et al. 10 Years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).
23. Franceschini, N. et al. Discovery and fine mapping of serum protein loci through transethnic meta-analysis. *Am. J. Hum. Genet.* **91**, 744–753 (2012).
24. Westra, H. et al. Cell specific eQTL analysis without sorting cells. *PLoS Genet.* **11**, e1005223 (2015).
25. Zhu, Z. et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**, 481–487 (2016).
26. Prins, B. et al. Investigating the causal relationship of C-reactive protein with 32 complex somatic and psychiatric outcomes: a large-scale cross-consortium mendelian randomization study. *PLoS Med.* **13**, e1001976 (2016).
27. Verduijn, M., Siegerink, B., Jager, K. J., Zoccali, C. & Dekker, F. W. Mendelian randomization: use of genetics to enable causal inference in observational studies. *Nephrol. Dial. Transplant.* **25**, 1394–1398 (2010).
28. Lawlor, D. A., Harbord, R. M., Sterne, J. A. C., Timpson, N. & Smith, G. D. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Stat. Med.* **27**, 1133–1163 (2008).
29. Hemani, G., Bowden, J. & Smith, G. D. Evaluating the potential role of pleiotropy in mendelian randomization studies. *Hum. Mol. Genet.* **27**, R195–R208 (2018).
30. Sullivan, P. F. et al. Psychiatric genomics: an update and an agenda. *Am. J. Psychiatry* **175**, 15–27 (2018).
31. Sekar, A. et al. Schizophrenia risk from complex variation of complement component 4. *Nature* **530**, 177–183 (2016).
32. Peterson, R. E. et al. The genetic architecture of major depressive disorder in Han Chinese women. *JAMA Psychiatry* **74**, 162–168 (2017).
33. Wood, A. R. et al. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.* **46**, 1173–1186 (2014).
34. Marouli, E. et al. Rare and low-frequency coding variants alter human adult height. *Nature* **542**, 186–190 (2017).
35. Yang, J. et al. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat. Genet.* **47**, 1114–1120 (2015).
36. Evangelou, E. et al. Genetic analysis of over 1 million people identifies 535 new loci associated with blood pressure traits. *Nat. Genet.* **50**, 1412 (2018).
37. Gibson, G. Rare and common variants: twenty arguments. *Nat. Rev. Genet.* **13**, 135–145 (2012).
38. Gratten, J., Wray, N. R., Keller, M. C. & Visscher, P. M. Large-scale genomics unveils the genetic architecture of psychiatric disorders. *Nat. Neurosci.* **17**, 782–790 (2014).
39. Uher, R. The role of genetic variation in the causation of mental illness: an evolution-informed framework. *Mol. Psychiatry* **14**, 1072–1082 (2009).
40. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An expanded view of complex traits: from polygenic to omnigenic. *Cell* **169**, 1177–1186 (2017).
41. Nolte, I. M. et al. Missing heritability: is the gap closing? An analysis of 32 complex traits in the lifelines cohort study. *Eur. J. Hum. Genet.* **25**, 877 (2017).
42. Culverhouse, R. C. et al. Collaborative meta-analysis finds no evidence of a strong interaction between stress and 5-HTTLPR genotype contributing to the development of depression. *Mol. Psychiatry* **23**, 133–142 (2018).
43. Gerritsen, L. et al. HPA axis genes, and their interaction with childhood maltreatment, are related to cortisol levels and stress-related phenotypes. *Neuropsychopharmacology* **42**, 2446–2455 (2017).
44. Peyrot, W. J. et al. Effects of polygenic risk scores on depression in childhood trauma. *Br. J. Psychiatry* **205**, 113–119 (2014).
45. Uher, R. & Zwicker, A. Etiology in psychiatry: embracing the reality of poly-gene-environmental causation of mental illness. *World Psychiatry* **16**, 121–129 (2017).
46. Ormel, J., Oldehinkel, A. J. & Brilman, E. I. The interplay and etiological continuity of neuroticism, difficulties and life events in the etiology of major and subsyndromal, first and recurrent depressive episodes in later life. *Am. J. Psychiatry* **158**, 885–891 (2001).
47. Caspi, A., Hariri, A. R., Holmes, A., Uher, R. & Moffitt, T. E. Genetic sensitivity to the environment: the case of the serotonin transporter gene and its implications for studying complex diseases and traits. *Am. J. Psychiatry* **167**, 509–527 (2010).
48. Weissman, M. M. et al. Cross-national epidemiology of major depression and bipolar disorder. *JAMA* **276**, 293–299 (1996).
49. Bifulco, A., Brown, G. W., Moran, P., Ball, C. & Campbell, C. Predicting depression in women: the role of past and present vulnerability. *Psychol. Med.* **28**, 39–50 (1998).
50. Brilman, E. I. & Ormel, J. Life events, difficulties and onset of depressive episodes in later life. *Psychol. Med.* **31**, 859–869 (2001).
51. Brown, G. W. Genetic and population perspectives on life events and depression. *Soc. Psychiatry Psychiatr. Epidemiol.* **33**, 363–372 (1998).
52. Gratten, J. & Visscher, P. M. Genetic pleiotropy in complex traits and diseases: implications for genomic medicine. *Genome Med.* **8**, 78 (2016).
53. Smoller, J. W. et al. Psychiatric genetics and the structure of psychopathology. *Mol. Psychiatry* **24**, 409–420 (2018).

54. Amare, A. T., Schubert, K. O., Klingler-Hoffmann, M., Cohen-Woods, S. & Baune, B. T. The genetic overlap between mood disorders and cardiometabolic diseases: a systematic review of genome wide and candidate gene studies. *Transl. Psychiatry* **7**, e1007 (2017).
55. Milaneschi, Y. et al. Genetic association of major depression with atypical features and obesity-related immunometabolic dysregulations. *JAMA Psychiatry* **74**, 1214–1225 (2017).
56. van den Oord, E. J. C. G. & Snieder, H. Including measured genotypes in statistical models to study the interplay of multiple factors affecting complex traits. *Behav. Genet.* **32**, 1–22 (2002).
57. Tropf, F. C. et al. Hidden heritability due to heterogeneity across seven populations. *Nat. Human. Behav.* **1**, 757–765 (2017).
58. Yang, J. et al. Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).
59. Polderman, T. J. C. et al. Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nat. Genet.* **47**, 702–709 (2015).
60. Cai, N., Kendler, K. & Flint, J. Minimal phenotyping yields GWAS hits of low specificity for major depression. *BioRxiv* https://doi.org/10.1101/440735 (2018).
61. Luciano, M. et al. Association analysis in over 329,000 individuals identifies 116 independent variants influencing neuroticism. *Nat. Genet.* **50**, 6–11 (2018).
62. Lamers, F. et al. Evidence for a differential role of HPA-axis function, inflammation and metabolic syndrome in melancholic versus atypical depression. *Mol. Psychiatry* **18**, 692–699 (2013).
63. Parker, G., Paterson, A. & Hadzi-Pavlovic, D. Cleaving depressive diseases from depressive disorders and non-clinical states. *Acta Psychiatr. Scand.* **131**, 426–433 (2015).
64. Wakefield, J. C. & Schmitz, M. F. Predictive validation of single-episode uncomplicated depression as a benign subtype of unipolar major depression. *Acta Psychiatr. Scand.* **129**, 445–457 (2014).
65. Nagel, M., Watanabe, K., Stringer, S., Posthuma, D. & van der Sluis, S. Item-level analyses reveal genetic heterogeneity in neuroticism. *Nat. Commun.* **9**, 905 (2018).
66. Ormel, J. et al. Neuroticism and common mental disorders: meaning and utility of a complex relationship. *Clin. Psychol. Rev.* **33**, 686–697 (2013).
67. Pearson, R. et al. Additive genetic contribution to symptom dimensions in major depressive disorder. *J. Abnorm. Psychol.* **125**, 495–501 (2016).
68. Kotov, R. et al. The hierarchical taxonomy of psychopathology (HiTOP): a dimensional alternative to traditional nosologies. *J. Abnorm. Psychol.* **126**, 454–477 (2017).
69. Waszczuk, M. A., Kotov, R., Ruggero, C., Gamez, W. & Watson, D. Hierarchical structure of emotional disorders: from individual symptoms to the spectrum. *J. Abnorm. Psychol.* **126**, 613–634 (2017).
70. Maier, R. M. et al. Improving genetic prediction by leveraging genetic correlations among human diseases and traits. *Nat. Commun.* **9**, 989 (2018).
71. Doudna, J. A. & Charpentier, E. The new frontier of genome engineering with CRISPR-Cas9. *Science* **346**, 1258096 (2014).
72. Zhang, F., Wen, Y. & Guo, X. CRISPR/Cas9 for genome editing: progress, implications and challenges. *Hum. Mol. Genet.* **23**, R40–R46 (2014).
73. Kathiresan, S. Developing medicines that mimic the natural successes of the human genome lessons from NPC1L1, HMGCR, PCSK9, APOC3, and CETP. *J. Am. Coll. Cardiol.* **65**, 1562–1566 (2015).
74. Cipriani, A. et al. Comparative efficacy and acceptability of 21 antidepressant drugs for the acute treatment of adults with major depressive disorder: a systematic review and network meta-analysis. *Lancet* **391**, 1357–1366 (2018).

The Novel *ASIC2* Locus Is Associated with Severe Gingival Inflammation

S. Zhang[1,2], K. Divaris[3,4], K. Moss[2], N. Yu[2], S. Barros[1,2], J. Marchesan[1,2], T. Morelli[1], C. Agler[5], S.J. Kim[2], D. Wu[1], K.E. North[4,6], J. Beck[2,7], and S. Offenbacher[1,2]

Abstract: *An increasing body of evidence suggests a significant genetic regulation of inflammatory response mechanisms; however, little is known regarding the genetic determinants of severe gingival inflammation (GI). We conducted a genome-wide association study of severe GI among 4,077 European American adults, participants in the Dental Atherosclerosis Risk in Communities cohort. The severe GI trait was defined dichotomously with the 90th percentile of gingival index ≥2 extent score. Genotyping was performed with the Affymetrix 6.0 array platform, and an imputed set of 2.5 million markers, based on HapMap Phase II CEU build 36, was interrogated. Genetic models were based on logistic regression and controlled for ancestry (10 principal components), sex, age, and examination center. One locus on chromosome 17 met genome-wide statistical significance criteria—lead single-nucleotide polymorphism: rs11652874 (minor allele frequency = 0.06, intronic to* ASIC2 *[acid-sensing ionic channel 2, formerly named* ACCN1*]; odds ratio = 2.1, 95% confidence interval = 1.6 to 2.7,* P *=* 3.9×10^{-8}*). This association persisted among subjects with severe periodontitis and was robust to adjustment for microbial plaque index. Moreover, the minor (G) allele was associated with higher levels of severe GI in stratified analyses among subsets of participants with high load of either "red" or "orange" complex pathogens, although this association was not statistically significant. While these results will require replication in independent samples and confirmation by mechanistic studies, this locus appears as a promising candidate for severe GI. Our findings suggest that genetic variation in* ASIC2 *is significantly associated with severe GI and that the association is plaque independent.*

Knowledge Transfer Statement: *Persistent gingival inflammation reflected by bleeding usually precedes ongoing attachment loss or periodontal disease progression. Our findings suggest that genetic variation in* ASIC2 *that is associated with severe gingival inflammation might be used as a genetic marker to identify people at higher risk for periodontal disease. Ongoing studies to uncover the mechanistic link between* ASIC2 *and gingival inflammation could lead to novel therapeutic interventions.*

Keywords: periodontal disease(s)/periodontitis, gingivitis, genetics, genomics, plaque/plaque biofilms, bacteria

Introduction

Periodontal disease is an abnormal inflammatory response to the pathogenic bacteria present in the biofilm. Gingival inflammation (GI), as determined by bleeding upon probing, is a clinical hallmark for both gingivitis and periodontitis. Gingivitis affects >80% of Americans, while 47% of American adults have periodontitis (Eke et al. 2012). In addition to stress

DOI: 10.1177/2380084416645290. [1]Department of Periodontology, School of Dentistry, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA; [2]Center for Oral and Systemic Diseases, School of Dentistry, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA; [3]Department of Pediatric Dentistry, School of Dentistry, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA; [4]Department of Epidemiology, UNC Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA; [5]Center for Oral and Craniofacial Health Sciences, School of Dentistry, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA; [6]Carolina Center for Genome Sciences, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA; [7]Department of Dental Ecology, School of Dentistry, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. Corresponding author: S. Offenbacher, 3501F Koury Oral Health Sciences Building, 385 S. Columbia Street, Chapel Hill, NC 27599-7455, USA. Email: Steven_Offenbacher@unc.edu

A supplemental appendix to this article is published electronically only at http://jdrctr.sagepub.com/supplemental.

exerted by the local microbial burden, it is now well-established that host genomics contributes to periodontal disease susceptibility. Twin studies by Michalowicz et al. (2000) suggested that heritability accounts for about 50% of the total phenotypic variability of periodontal disease. More recently, in a study of >10,000 Swedish twin pairs, Mucci et al. (2005) reported heritability estimates of 39% and 33% for women and men, respectively. In a more recent report, our group used a genome-wide association (GWA) approach to show that approximately 22% of phenotypic variance observed in severe chronic periodontitis can be explained by a GWA study (GWAS) set of approximately 650,000 genotyped single-nucleotide polymorphisms (SNPs) with minor allele frequency (MAF) ≥5% (Divaris et al. 2013).

Similar to other complex trait diseases, predisposition to periodontal disease is likely polygenic (Kinane et al. 2005). Most previous studies interrogating the genetic role in periodontal disease were based on a candidate-gene approach (Kornman et al. 1997; Tervonen et al. 2007; Kobayashi et al. 2009; Laine et al. 2012). GWASs provide an unbiased approach for generating hypotheses on nonrare genetic variants for common diseases, such as periodontitis. By the end of 2013, approximately 14,000 SNPs had reported GWAs with >600 traits (Welter et al. 2014). Although SNPs identified by GWAS may explain a small proportion of the observed variance for most diseases (Gibson 2010), this methodology has brought about notable successes in the study of obesity and metabolic traits (Locke et al. 2015). There is promise that similar advancements will eventually become possible in oral and periodontal research.

There are only a few GWASs on chronic periodontitis (Divaris et al. 2013; Teumer et al. 2013; Feng et al. 2014; Shaffer et al. 2014; Shimizu et al. 2015). While no significant genome-wide SNP hits have been identified so far, several suggestive loci have been reported. Divaris and colleagues (2013) reported the *NIN* locus as showing suggestive evidence of association with severe periodontitis. This locus was highlighted in a separate GWAS of periodontal disease by Shaffer et al. (2014), although the top SNP in the latter study was approximately 350 kb upstream of *NIN* and not in linkage disequilibrium with the lead SNP reported by Divaris et al. In addition, the *NPY* locus, which also emerged from the Atherosclerosis Risk in Communities (ARIC) study report, was subsequently highlighted by Freitag-Wolf et al. (2014) as being associated with periodontitis among men (albeit tagging SNPs were not in linkage disequilibrium), as well as by a gene-centric reanalysis of ARIC reported by Rhodin et al. (2014). Even fewer GWASs have explored loci associated with periodontal phenotypes defined by quantitative clinical signs rather than disease categories. Two notable exceptions include a study by Shaffer et al. (2014) that used probing depth ≥5.5 mm present in at least 2 sextants as the outcome for their GWA analyses and a study by Divaris et al. (2012) that reported several suggestive genetic loci associated with high bacterial pathogen colonization by "red" complex (*Porphyromonas gingivalis, Tannerella Forsythia*, and *Treponema denticola*), "orange" complex (*Prevotella intermedia, Fusobacterium nucleatum, Prevotella nigrescens*, etc.), and *Aggregatibacter actinomycetemcomitans* among European American participants of the ARIC cohort.

There is compelling evidence that persistent GI reflected by bleeding upon probing combined with residual probing depth after periodontal treatment has prognostic value for ongoing attachment loss or disease progression (Lang et al. 1986; Claffey et al. 1990; Matuliene et al. 2008). GI may also contribute to a systemic hyperinflammatory state that is etiologically related to other chronic diseases and pathologies (Scannapieco 2004). We do know that the microbial biofilm is etiologically associated with gingivitis (Loe et al. 1965). Yet, it is also known that inflammation, including the production of cytokines, is highly genetically regulated (de Craen et al. 2005). However, no systematic effort has been undertaken to study the genetic underpinning of GI as an independent clinical sign with a GWA approach. It is also unknown whether a genetic association, if present, is modified by the level of exposure to periodontal pathogens. We hypothesize that genetic variants that interact with known periodontal pathogens are associated with the severe GI trait as reflected by bleeding upon probing. To test this hypothesis, we used the well-defined Dental ARIC cohort and conducted the first GWA scan for genetic risk loci associated with extensive severe GI and examined these associations among subgroups with varying levels of subgingival periodontal pathogens and categorical classifications of chronic periodontitis.

Materials and Methods

Study Population and Measurements

We used data from the previously reported GWAS performed in a subset of European American subjects who were participants in the ARIC study (ARIC Investigators 1989). The ARIC cohort study aimed to study atherosclerosis, cardiovascular disease risk factors, and outcomes and originally recruited 15,792 community-dwelling residents in 4 U.S. communities between 1987 and 1989. As an ancillary study, Dental ARIC recruited a subset of ARIC participants in the fourth ARIC visit (1996-1998). During their dental visit, 6,017 dentate ARIC subjects had a complete full-mouth dental examination, which included measurements of probing depth, attachment loss, and bleeding upon probing at 6 sites per tooth, including third molars, number of missing teeth, and gingival index (Loe and Silness, 1963). Plaque index (PI) was recorded for each tooth according to the method introduced by Silness and Loe (1964). Subgingival plaque samples were also collected from 973 European American Dental ARIC participants. For the present analyses, we included only European

American participants and excluded individuals whose genotyping did not meet quality control criteria. Therefore, in this report we used genotype and clinical data from a subset of 4,077 Dental ARIC subjects.

Genotyping, Quality Control, and Imputation

Detailed description of genotyping and data processing were described elsewhere (Divaris et al. 2012; Divaris et al. 2013). Briefly, DNA was extracted from blood samples. Genotyping was performed with the Affymetrix Genome-Wide Human SNP Array 6.0 chip containing 906,600 SNP markers. After quality control, an imputation to 2.5 million SNP markers was performed with 669,450 high-quality genotyped SNPs and the MACH program (version 1.0.16) based on HapMap Phase II CEU build 36. Further SNP exclusion included the following criteria: imputation quality score <0.8, missing data rate >10% after imputation, and MAF <5%.

Quantification of Periodontal Organisms in Plaque Samples by DNA Checkerboard

Levels of bacteria within plaque samples were determined as previously described (Socransky et al. 1994). One plaque sample taken from the subgingival mesiobuccal site of the maxillary right first molar was used from each subject and hybridized to a DNA chromosomal checkerboard for the 8 periodontal pathogens present in the red or orange complexes defined by Socransky et al. (1998). Organism levels were expressed as counts based on known microbial standards. A "high" load was defined as those in the upper quartile (25%) range of counts, and a "low" load, those in the lowest three-quartiles count range. Eight periodontal pathogens were included for analyses: *Porphyromonas gingivalis, Prevotella intermedia, Treponema denticola, Tannerella forsythia* (formerly *Bacteroides forsythus*), *Campylobacter rectus, Fusobacterium nucleatum, Aggregatibacter actinomycetemcomitans* (formerly *Actinobacillus actinomycetemcomitans*), and *Prevotella nigrescens.*

Analytic Strategy

The clinical trait of interest was the extent of moderate to severe GI as reflected by the ordinal gingival index, wherein "0" reflects normal gingiva; "1," mild inflammation but no bleeding upon running probe in the sulcus; "2," moderate inflammation with bleeding upon running probe in the sulcus; and "3," severe inflammation with a tendency for spontaneous bleeding. Extent scores were calculated as the percentage of sites exhibiting gingival index ≥2 (EGIGE2) among all examined sites for each participant. The severe GI trait was defined dichotomously as gingival index extent score in the 90th percentile (top 10%), whereas the continuous GI variable was used for exploratory and stratified analyses. Genetic models were based on logistic regression, adjusting for ancestry (10 principal components), age, sex, and examination center. A conventional genome-wide significance level of $P < 5 \times 10^{-8}$ was used for single-marker discovery in the GWAS, whereas loci with suggestive evidence of association ($P < 5 \times 10^{-6}$) are also reported. All genetic analyses were performed with the ProbABEL software. We utilized Manhattan and LocusZoom plots (Pruim et al. 2010) to visualize the genetic loci of interest. We followed HUGO Gene Nomenclature to report genes. Exploratory analyses of the continuous GI trait, as well as analyses stratified by chronic periodontitis diagnosis and high versus low pathogen load, were based on generalized linear models, adjusting for ancestry, age, sex, examination center, and microbial plaque levels and accounting for multiple comparisons. We followed the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) checklist guidelines for the reporting of findings.

Results

Single-Marker GWAS

Genomic inflation for the severe GI association analysis was low ($\lambda = 1.008$). The quantile-quantile plot (Appendix Fig.) did not reveal any evidence of residual population stratification. The full list of GWAS results, including estimates of association for all 2.5 million SNPs, is available at http://genomewide.net/public/aric/dental/gingivitis/EGIGE2.txt. One locus on chromosome 17q11.2 met genome-wide statistical significance criteria (Fig. 1A) for the severe GI trait. The lead SNP in this locus was rs11652874: MAF (G) = 0.06, odds ratio (OR) = 2.1, 95% confidence interval = 1.6 to 2.7, $P = 3.9 \times 10^{-8}$. This imputed SNP has an imputation quality score of 0.998, which is considered excellent. This SNP is intronic to the *ASIC2* gene (acid-sensing ionic channel 2; previously named *ACCN1*; Fig. 1B), within the intron region between exons 2 and 3, and in tight linkage disequilibrium with 9 additional intronic SNPs (Fig. 1C). We found no association of this SNP with clinical diagnoses of chronic periodontitis (e.g., 4-level case definition per the Centers for Disease Control and Prevention / American Academy of Periodontology [CDC/AAP]; $P = 0.14$; Table). We found 3 additional loci with suggestive evidence of association ($P < 5 \times 10^{-6}$) with severe GI: *PHGDH* (lead SNP rs894049, $P = 1.2 \times 10^{-6}$, OR = 1.62, MAF [G] = 0.14), *NRXN1* (lead SNP rs1520455, OR = 1.47, $P = 4.0 \times 10^{-6}$, MAF [C] = 0.37), and *ACVR1* (lead SNP rs7578180, OR = 1.43, $P = 4.5 \times 10^{-6}$, MAF [C] = 0.33). *ASIC2* encodes a member of the degenerin/epithelial sodium channel superfamily and is likely involved in neurotransmission (Krishtal 2003).

Stratified Analyses

To further explore the *ASIC2* locus signal, we examined the lead SNP's association with the continuous GI trait (EGIGE2) stratified by the 4-level CDC/AAP chronic periodontitis diagnoses (Fig. 2). Not surprisingly, in both *ASIC2* genotype groups, EGIGE2 was higher among participants with more advanced (severe) disease status versus those with milder disease. We found no association between rs11652874 and EGIGE2 among

Figure 1. rs11652874 in the *ASIC2* gene locus is associated with severe gingival inflammation. (**A**) Manhattan plot of genome-wide association analysis results of severe gingival inflammation (binary trait based on EGIGE2q90) in the Atherosclerosis Risk in Communities cohort ($n = 4,077$). (**B**) Locus zoom visualization of the 17q11.2 locus (*ASIC2*, lead SNP: rs11652874) flanking ±400 kb. The vertical axis represents the $\log_{10}$-transformed *P* value for each SNP. A schematic representation of the *ASIC2* gene exons in relation to the lead SNP (rs11652874) is shown in the inset. (**C**) Locus zoom visualization of the 17q11.2 locus flanking a zoomed-in ±25-kb region. SNP, single-nucleotide polymorphism.

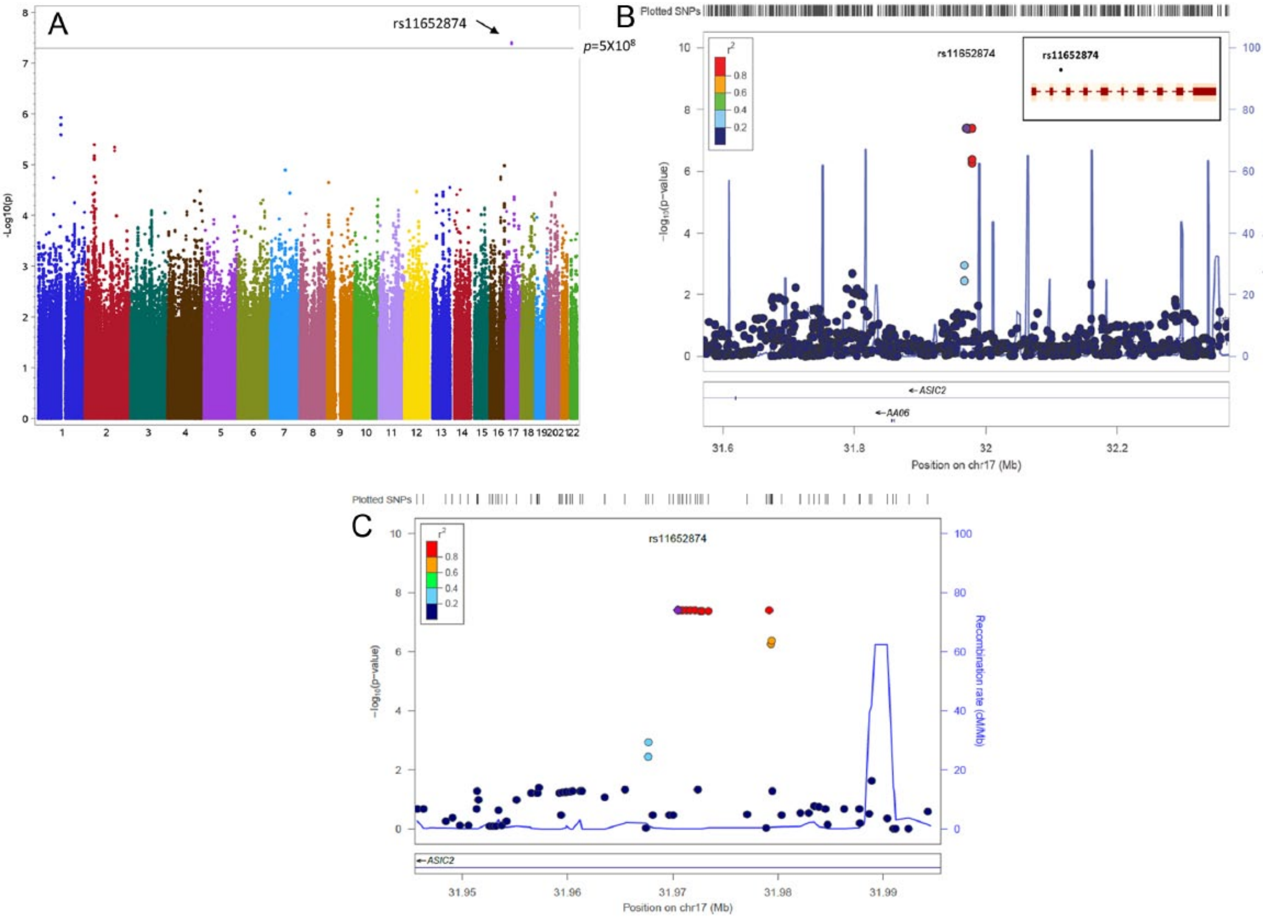


Table.

Genotype Distribution of rs11652874 according to Periodontal Disease Definitions.

Genotype	Periodontal Disease Definition, *n* (%)[a]				
	Healthy	Mild	Moderate	Severe	Total
C/C	492 (10.3; 11.7)	1,178 (24.7; 28.0)	1,817 (38.1; 43.2)	720 (15.1; 17.1)	4,207 (88.3; 100.0)
C/G+G/G[b]	67 (1.4; 12.0)	177 (3.7; 31.7)	214 (4.5; 38.3)	101 (2.1; 12.3)	559 (11.7; 100.0)
Total	559 (11.7)	1,355 (28.5)	2,031 (42.6)	821 (17.2)	4,766 (100.0)

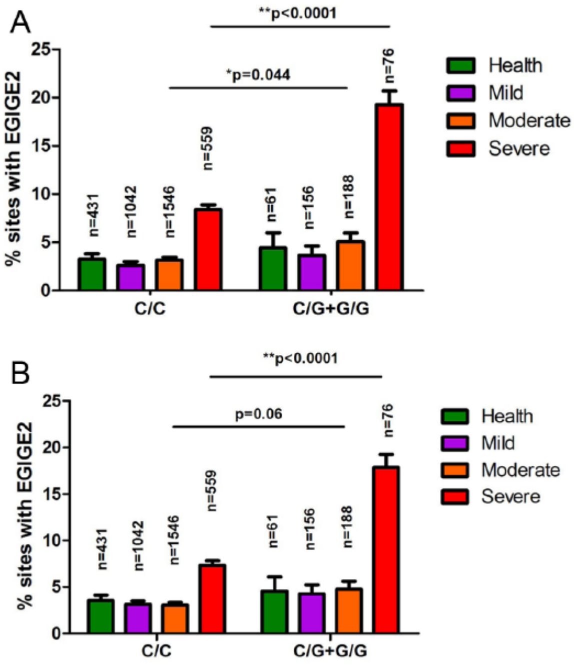
Four-level case definition based on the American Academy of Periodontology and Centers for Disease Control and Prevention (Eke et al. 2012).

[a] For cells that have 2 values in parentheses, the first value denotes the percentage among the entire sample, and the second, the percentage among subjects with the same genotype.

[b] $P = 0.14$ (χ^2 test).

subjects with periodontal health ($P = 0.67$) or mild periodontitis ($P = 0.13$). However, among subjects with moderate or severe periodontitis, the EGIGE2 is significantly higher among rs11652874 (G) allele carriers versus noncarriers ($P = 0.04$ for moderate periodontitis and $P < 0.0001$ for severe periodontitis) after adjustment for study design features,

Figure 2. Association of rs11652874 polymorphisms (minor allele [G] carriers vs. noncarriers) with gingival inflammation extent score ≥2 (EGIGE2) stratified by chronic periodontitis diagnosis (case status classification per the American Academy of Periodontology and Centers for Disease Control and Prevention), adjusting for (**A**) age, sex, ancestry (10 principal components), and examination center and (**B**) microbial plaque levels, age, sex, ancestry (10 principal components), and examination center.



participant characteristics, and ancestry (Fig. 2A). This association was somewhat attenuated after further adjustment for plaque score (Fig. 2B) in the moderate periodontitis stratum (unadjusted: beta = 1.92, $P = 0.04$, vs. adjusted: beta = 1.69, $P = 0.06$), whereas it remained significant ($P < 0.0001$) in the severe periodontitis stratum. In other words, the increased GI among those with the *ASIC2* minor (G) allele polymorphism was not substantially affected by microbial plaque levels, particularly among participants with severe chronic periodontitis.

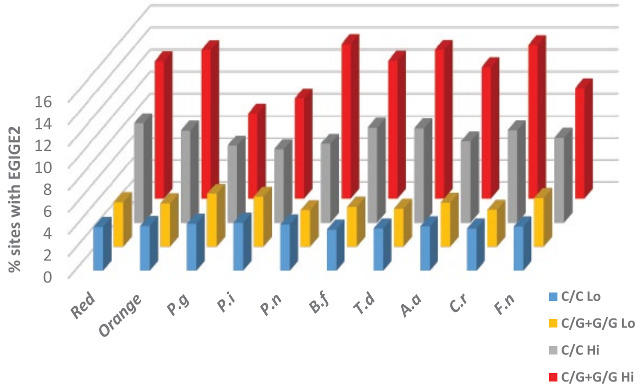
To further explore the possible effect modification by microbial plaque, next we examined possible SNP × periodontal pathogen interactions that might account for the higher gingival index score. Therefore, we examined whether rs11652874 (G) allele carrier status modified the gingival inflammatory response to bacteria of the red and the orange complexes, as well as individual bacterial species. High periodontal pathogen loads were associated with statistically increased gingival index scores as compared with lower bacterial loads, independent of rs1652874 genotype (Fig. 3; *P* values not shown). Although a trend for an association between rs1652874 (G) allele and periodontal pathogen levels was noted, we found no statistically significant association after adjustment for multiple comparisons.

Discussion

In this genome-wide investigation conducted among the well-defined Dental ARIC cohort, we found a novel locus associated with severe GI. The identified polymorphism is intronic to the *ASIC2* gene and not associated with CDC/AAP chronic periodontitis diagnosis. The rs11652874 minor (G) allele was associated with the most severe gingival bleeding among those with severe chronic periodontitis, and this was independent of microbial plaque levels. We tested the hypothesis that levels of specific periodontal microorganisms may alter the clinical response according to the *ASIC2* polymorphism. Although some trends were noted, the extent of GI was not associated with this polymorphism within red or orange pathogen high- and low-load groups. These findings suggest that there may be additional loci or biological or environmental/behavioral risk factors that are specific to severe disease and interact to produce extremely high gingival index scores (Manolio et al. 2009).

The lead SNP in the *ASCI2* locus is located in the gene's second intron of the gene. Acid-sensing ion channels play an important role in the nervous system (Lingueglia and Lazdunski 2015). We found no important functional annotation or expression quantitative trait loci information about this SNP in publicly available databases such as SNPnexus (http://snp-nexus.org/) and GTEx (http://www.gtexportal.org/). *ASIC2* has been reported to be ubiquitously expressed in the mammalian peripheral and central nervous system. Residing in the free sensory nerve endings that are in contact with local tissue, this molecule may regulate the physiologic response of microvessels to fluctuations of local oxygen consumption. During inflammation, the drop of pH due to increased oxygen utilization leads to *ASIC2* activation and the release of vasoactive substances from the other branches of nerve endings is triggered (Krishtal 2003). The involvement of *ASIC2* in inflammation is supported by experiments with *ASIC2* knockout mice, in which several inflammatory markers, such as TGF-β (transforming growth factor-β) and TAPA-1 (target of antiproliferative antibody 1), were all upregulated when compared with wild-type controls (Gannon et al. 2015). Although there is evidence supporting a plausible role for *ASIC2* contributing to gingival hyperinflammatory phenotype, further replication and mechanistic studies are warranted for its validation.

Figure 3. Comparison of severe gingival inflammation extent scores (EGIGE2) by levels of 8 periodontal pathogens in the red and orange complexes as determined by DNA checkerboard; bacterial loads were dichotomized into high level (the top quartile) and low level (lowest 3 quartiles) and stratified by genotypes at rs11652874. *A.a, Aggregatibacter actinomycetemcomitans* (formerly *Actinobacillus actinomycetemcomitans*); *B.f, Tannerella forsythia* (formerly *Bacteriodes forsythus*); *C.r, Campylobacter rectus; F.n, Fusobacterium nucleatum; P.g, Porphorymonas gingivalis; P.i, Prevotella intermedia; P.n, Prevotella nigrescens; T.d, Treponema denticola*.



The neurotransmitter and nervous system signaling pathways have been repeatedly reported in our GWASs and others, as well as in our transcriptome studies (Offenbacher et al. 2009; Divaris et al. 2013). The exact role of neuropathways in periodontal health and disease remains to be elucidated; nevertheless, neuropeptides are known to regulate acid-sensing ion channels, as detailed by Vick and Askwith (2015). Ingenuity pathway analyses based on data from our experimental transcriptome gingivitis study indicate that there are 4 genes upstream to *ASIC2* that are upregulated during gingivitis induction (*ASIC2* was not present on our transcriptome array for scanning), suggesting the activation of inflammatory pathways involving *ASIC2*. These genes are *NRXN1, CHD8, CTNNB1*, and *NLGN1* (Offenbacher et al. 2009). Noteworthy, several chemokine genes, including *CCL2, CCL8, CCL11, CCL7*, and *CCL13*, are located downstream of *ASCI2* on chromosome 17q12. The involvement of those CCL chemokines in periodontal inflammation is well documented (Silva et al. 2007). The signal in the *ASIC2* locus may also highlight other possibly associated genes in nearby genome regions, and these chemokine genes would serve as excellent additional candidates.

In addition to *ASIC2*, we identified several suggestive loci close to genes with known functions. For example, rs894079 is located in the 5′ untranslated region of *PHGDH*, which encodes the enzyme involved in the early steps of L-serine synthesis in animal cells (Zogg 2014). Rs1520455 is intronic to *NRXN1*, which encodes neurexin, a presynaptic protein diffusely distributed in neurons and a key component in neuronal network (Bottos et al. 2011). Rs7578180 is intronic to *ACVR1*, which encodes a receptor for activins, which are dimeric growth and differentiation factors belonging to the TGF-β superfamily of signaling proteins (Goumans et al. 2003). Although those SNPs did not meet criteria for genome-wide statistical significance, they appear as promising candidates for future studies.

Our study is limited by the absence of a replication sample to externally validate the observed association between rs11652874 and GI. Although our results are derived from a moderately sized GWAS, including >4,000 community-dwelling participants in a well-defined cohort, an independent replication of this SNP and other polymorphisms highlighted in this study is warranted. We acknowledge that independent replication and mechanistic validation are crucial because this association—although statistically robust—may simply represent a false-positive finding (Ioannidis 2005). Moreover, we considered levels of only 8 key pathogens within the red and orange complexes and cannot exclude that other organisms unique to the moderate and severe disease states might be critical and interact with this polymorphism to produce high levels of GI. For example, we and others have reported on the importance of organisms such as *Fretibacterium* and *Treponema* species in periodontal disease (Marchesan et al. 2015). This recent Human Microbiome Identification Microarray study identified 272 subgingival microbial species and defined *Synergistetes*- and *Spirochaetes*-dominated microbial community structures that were significantly associated with increased probing depth and bleeding upon probing (Marchesan et al. 2015). Those pathogens and many other important organisms were not included in analyses presented here, which were limited to the key "classical" cultivable pathogens.

In summary, in these GWA analyses of severe GI among the Dental ARIC European American population we found a genome-wide significant association of a chromosome 17q12 locus, marked by a SNP intronic to the *ASIC2* gene. The association persisted among groups with severe periodontitis and was independent of clinical levels of microbial plaque; neither loads of specific red or orange complex pathogens modified this association. These results provide support for the conduct of further studies in this or nearby genome regions to explore if and how common genetic variation in *ASIC2* plays a role in severe GI.

Author Contributions

S. Zhang, K. Divaris, contributed to conception, data acquisition, analysis, and interpretation, drafted and critically revised the manuscript; K. Moss, contributed to data acquisition and analysis, drafted the manuscript; N. Yu, contributed to data acquisition and interpretation, drafted and critically revised the manuscript; S. Barros, T. Morelli, S.J. Kim, D. Wu, contributed to data interpretation, critically revised the manuscript; J. Marchesan, contributed to data acquisition and analysis, critically revised the manuscript; C. Agler, contributed to data analysis, drafted the manuscript; K.E. North, contributed to data analysis and interpretation, drafted and critically revised the manuscript; J. Beck, contributed to conception, critically revised the manuscript; S. Offenbacher, contributed to conception, design, data acquisition, analysis, and interpretation, drafted and critically revised the manuscript. All authors gave final approval and agree to be accountable for all aspects of the work.

Acknowledgments

We thank the staff and participants of the ARIC study for their important contributions. The ARIC study is carried out as a collaborative study supported by the National Heart, Lung, and Blood Institute contracts (HHSN268201100005C, HHSN268201100006C, HHSN268201100007C, HHSN268201100008C, HHSN268201100009C, HHSN268201100010C, HHSN268201100011C, HSN268201100012C) and grants (R01HL087641, R01HL59367, R01HL086694), the National Human Genome Research Institute (contract U01HG004402), the National Institutes of Health (contract HHSN268200625226C), the National Institute of Environmental Health Sciences (grant P30ES010126), and the National Institute of Dental and Craniofacial Research (grants R01DE11551, R01DE021418). Infrastructure was partly supported by a component of the National Institutes of Health and NIH Roadmap for Medical Research (grant UL1RR025005). This study was funded also by National Institutes of Health training (grant R90DE022527; to S.Z.). The authors declare no potential conflicts of interest with respect to the authorship and/or publication of this article.

References

ARIC Investigators. 1989. The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. Am J Epidemiol. 129(4):687–702.

Bottos A, Rissone A, Bussolino F, Arese M. 2011. Neurexins and neuroligins: synapses look out of the nervous system. Cell Mol Life Sci. 68(16):2655–2666.

Claffey N, Nylund K, Kiger R, Garrett S, Egelberg J. 1990. Diagnostic predictability of scores of plaque, bleeding, suppuration and probing depth for probing attachment loss: 3 1/2 years of observation following initial periodontal therapy. J Clin Periodontol. 17(2):108–114.

de Craen AJ, Posthuma D, Remarque EJ, van den Biggelaar AH, Westendorp RG, Boomsma DI. 2005. Heritability estimates of innate immunity: an extended twin study. Genes Immun. 6(2):167–170.

Divaris K, Monda KL, North KE, Olshan AF, Lange EM, Moss K, Barros SP, Beck JD, Offenbacher S. 2012. Genome-wide association study of periodontal pathogen colonization. J Dent Res. 91(7):21S-28S.

Divaris K, Monda KL, North KE, Olshan AF, Reynolds LM, Hsueh WC, Lange EM, Moss K, Barros SP, Weyant RJ, et al. 2013. Exploring the genetic basis of chronic periodontitis: a genome-wide association study. Hum Mol Genet. 22(11):2312–2324.

Eke PI, Dye BA, Wei L, Thornton-Evans GO, Genco RJ; CDC Periodontal Disease Surveillance Workgroup. 2012. Prevalence of periodontitis in adults in the United States: 2009 and 2010. J Dent Res. 91(10):914–920.

Feng P, Wang X, Casado PL, Kuchler EC, Deeley K, Noel J, Kimm H, Kim JH, Haas AN, Quinelato V, et al. 2014. Genome wide association scan for chronic periodontitis implicates novel locus. BMC Oral Health. 14:84.

Freitag-Wolf S, Dommisch H, Graetz C, Jockel-Schneider Y, Harks I, Staufenbiel I, Meyle J, Eickholz P, Noack B, Bruckmann C, et al. 2014. Genome-wide exploration identifies sex-specific genetic effects of alleles upstream npy to increase the risk of severe periodontitis in men. J Clin Periodontol. 41(12):1115–1121.

Gannon KP, McKey SE, Stec DE, Drummond HA. 2015. Altered myogenic vasoconstriction and regulation of whole kidney blood flow in the asic2 knockout mouse. Am J Physiol Renal Physiol. 308(4):F339–F348.

Gibson G. 2010. Hints of hidden heritability in GWAS. Nat Genet. 42(7):558–560.

Goumans MJ, Lebrin F, Valdimarsdottir G. 2003. Controlling the angiogenic switch: a balance between two distinct TGF-b receptor signaling pathways. Trends Cardiovas Med. 13(7):301–307.

Ioannidis JP. 2005. Why most published research findings are false. PLoS Med. 2(8):e124.

Kinane DF, Shiba H, Hart TC. 2005. The genetic basis of periodontitis. Periodontology 2000. 39:91–117.

Kobayashi T, Murasawa A, Ito S, Yamamoto K, Komatsu Y, Abe A, Sumida T, Yoshie H. 2009. Cytokine gene polymorphisms associated with rheumatoid arthritis and periodontitis in japanese adults. J Periodontol. 80(5):792–799.

Kornman KS, Crane A, Wang HY, di Giovine FS, Newman MG, Pirk FW, Wilson TG Jr., Higginbottom FL, Duff GW. 1997. The interleukin-1 genotype as a severity factor in adult periodontal disease. J Clin Periodontol. 24(1):72–77.

Krishtal O. 2003. The ASICs: signaling molecules? Modulators? Trends Neurosci. 26(9):477–483.

Laine ML, Crielaard W, Loos BG. 2012. Genetic susceptibility to periodontitis. Periodontology 2000. 58(1):37–68.

Lang NP, Joss A, Orsanic T, Gusberti FA, Siegrist BE. 1986. Bleeding on probing: a predictor for the progression of periodontal disease? J Clin Periodontol. 13(6):590–596.

Lingueglia E, Lazdunski M. 2015. Acid-sensing ion channels in the nervous system: foreword. Neuropharmacology. 94:1.

Locke AE, Kahali B, Berndt SI, Justice AE, Pers TH, Day FR, Powell C, Vedantam S, Buchkovich ML, Yang J, et al. 2015. Genetic studies of body mass index yield new insights for obesity biology. Nature. 518(7538):197–206.

Loe H, Silness J. 1963. Periodontal disease in pregnancy: I. Prevalence and severity. Acta Odontol Scand. 21:533–551.

Loe H, Theilade E, Jensen SB. 1965. Experimental gingivitis in man. J Periodontol. 36:177–187.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, et al. 2009. Finding the missing heritability of complex diseases. Nature. 461(7265):747–753.

Matuliene G, Pjetursson BE, Salvi GE, Schmidlin K, Bragger U, Zwahlen M, Lang NP. 2008. Influence of residual pockets on progression of periodontitis and tooth loss: results after 11 years of maintenance. J Clin Periodontol. 35(8):685–695.

Marchesan JT, Morelli T, Moss K, Barros SP, Ward M, Jenkins W, Aspiras MB, Offenbacher S. 2015. Association of synergistetes and cyclodipeptides with periodontitis. J Dent Res. 94(10):1425–1431.

Michalowicz BS, Diehl SR, Gunsolley JC, Sparks BS, Brooks CN, Koertge TE, Califano JV, Burmeister JA, Schenkein HA. 2000. Evidence of a substantial genetic basis for risk of adult periodontitis. J Periodontol. 71(11):1699–1707.

Mucci LA, Bjorkman L, Douglass CW, Pedersen NL. 2005. Environmental and heritable factors in the etiology of oral diseases: a population-based study of Swedish twins. J Dent Res. 84(9):800–805.

Offenbacher S, Barros SP, Paquette DW, Winston JL, Biesbrock AR, Thomason RG, Gibb RD, Fulmer AW, Tiesman JP, Juhlin KD, et al. 2009. Gingival transcriptome patterns during induction and resolution of experimental gingivitis in humans. J Periodontol. 80(12):1963–1982.

Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP, Boehnke M, Abecasis GR, Willer CJ. 2010. Locuszoom: regional visualization of genome-wide association scan results. Bioinformatics. 26(18):2336–2337.

Rhodin K, Divaris K, North KE, Barros SP, Moss K, Beck JD, Offenbacher S. 2014. Chronic periodontitis genome-wide association studies: gene-centric and gene set enrichment analyses. J Dent Res. 93(9):882–890.

Scannapieco FA. 2004. Periodontal inflammation: From gingivitis to systemic disease? Compend Contin Educ Dent. 25(7 Suppl 1):16–25.

Shaffer JR, Polk DE, Wang X, Feingold E, Weeks DE, Lee MK, Cuenco KT, Weyant RJ, Crout RJ, McNeil DW, et al. 2014. Genome-wide association study of periodontal health measured by probing depth in adults ages 18–49 years. G3 (Bethesda). 4(2):307–314.

Shimizu S, Momozawa Y, Takahashi A, Nagasawa T, Ashikawa K, Terada Y, Izumi Y, Kobayashi H, Tsuji M, Kubo M, et al. 2015. A genome-wide association study of periodontitis in a japanese population. J Dent Res. 94(4):555–561.

Silness J, Loe H. 1964. Periodontal disease in pregnancy: II. Correlation between oral hygiene and periodontal condtion. Acta Odontol Scand. 22:121–135.

Silva TA, Garlet GP, Fukada SY, Silva JS, Cunha FQ. 2007. Chemokines in oral inflammatory diseases: apical periodontitis and periodontal disease. J Dent Res. 86(4):306–319.

Socransky SS, Haffajee AD, Cugini MA, Smith C, Kent RL Jr. 1998. Microbial complexes in subgingival plaque. J Clin Periodontol. 25(2):134–144.

Socransky SS, Smith C, Martin L, Paster BJ, Dewhirst FE, Levin AE. 1994. “Checkerboard” DNA-DNA hybridization. BioTechniques. 17(4):788–792.

Tervonen T, Raunio T, Knuuttila M, Karttunen R. 2007. Polymorphisms in the CD14 and IL-6 genes associated with periodontal disease. J Clin Periodontol. 34(5):377–383.

Teumer A, Holtfreter B, Volker U, Petersmann A, Nauck M, Biffar R, Volzke H, Kroemer HK, Meisel P, Homuth G, et al. 2013. Genome-wide association study of chronic periodontitis in a general german population. J Clin Periodontol. 40(11):977–985.

Vick JS, Askwith CC. 2015. Asics and neuropeptides. Neuropharmacology. 94:36–41.

Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L, et al. 2014. The NHGRI GWAS catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 42(Database issue):D1001–D1006.

Zogg CK. 2014. Phosphoglycerate dehydrogenase: potential therapeutic target and putative metabolic oncogene. J Oncol. 2014:524101.

genes

Editorial

Genetics and Genomics of Addiction

Jason A. Bubier

The Jackson Laboratory, Bar Harbor, ME 04609, USA; jason.bubier@jax.org

Substance use disorders (SUD), like many neuropsychiatric conditions, are a heterogeneous group of disorders with similar symptomatology but often different pathoetiology. As diagnosed by the DSM-V, SUD requires at least two of eleven diagnostic criteria, resulting in the possibility of over 2000 different combinations of symptoms. Further segmenting the disorder is that SUD can involve a variety of legal and illegal substances. There are currently approaches to treating and managing SUD, but we have a limited armamentarium. Genetic and genomic studies of addiction can facilitate a more accurate molecular diagnosis and, ultimately, the development of appropriate treatments. Sequenced genomes, new animal model resources, and Genome Wide Association Studies (GWAS) of large sample sizes have made vast inroads into our understanding of the genetics that underlies addiction. Future mechanistic studies will be driven by integrating big data from animal models and human studies to identify their consilient features. This Special Issue, "Genetics and Genomics of Addiction," focuses on genetic contributions to this disease that may lead to better-targeted therapeutics and diagnostics. This issue contains five original research articles and one review paper that furthers our collective knowledge of SUD disease etiology and the genetic risk factors underlying the disease.

Chunduri, Watson and Ashbrook explore how the controlled laboratory environments of animal model research can be leveraged to understand the gene-by-environment (GxE) interactions underlying SUD [1]. Their work relies on the recombinant inbred population of the BXD strains of mice, which have reproducible genomes that can be tested in different laboratory environments, allowing one to make novel insights into GxE. They illustrate that this can be accomplished using GeneNetwork.org, a database and tool suite containing over 10,000 phenotypes and over 100 'omics datasets measured under various environmental conditions derived from the BXD panel. The group reanalyzed data published in 2010 using new mapping software and complete genomic sequence and identified novel candidate genes involved in response to drug use. Also, using model organisms, Rizk et al. [2] explored the impact of sex and circadian disruption on alcohol drinking in mice. Utilizing both sexes of *Clock* mutants, the researchers assessed alcohol consumption and preference. As expected, female control inbred C57BL/6J mice drank more than the males. However, there were no sex differences in the *Clock* mutants, which displayed greater alcohol preference and consumption over controls of both sexes. Studies like these illustrate how the environment and circadian disruption can affect SUD. The final article in this collection utilizing animal models was the work of Lie et al. [3] in which the authors set out to determine the relationship between methylation of the guanine nucleotide-binding protein, α stimulating complex locus, (*Gnas)* promoter, *Gnas* mRNA levels and morphine reward memory. Hypermethylation of the *Gnas* promoter and decreased mRNA expression in the basolateral amygdala were necessary for reconsolidating opiate-associate memories. A review paper by Sambo and Goldman [4] summarizes the current knowledge around the role of genetic influence in Fetal Alcohol Spectrum Disorder (FASD). This SUD-adjacent field of study explores the range of outcomes that result from prenatal alcohol exposure. The human literature related to twin studies, copy number variations, and candidate gene studies, illustrates a role for various genetic factors in FASD susceptibility. Fedorenko et al. [5] also explored alcohol use, assessing the association of phosphatidylinositol-5-phosphate-4-kinase type two α (PIP4K2A) polymorphisms with alcohol use disorder. Polymorphisms

Citation: Bubier, J.A. Genetics and Genomics of Addiction. *Genes* **2023**, *14*, 1760. https://doi.org/10.3390/genes14091760

Received: 11 June 2023
Accepted: 16 June 2023
Published: 4 September 2023

Copyright: © 2023 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

in PIP4K2a, a gene known to play a role in dopaminergic pathways, were tested for their association with AUD in a population of Russian men. In this case-control study, two PIP4K2A SNPs had a higher association of occurring in cases than in controls, suggesting that different PIP4K2A polymorphisms may be associated with an increased risk of AUD. The final paper in this collection, that of Chmielowiec et al. [6], also performed case-control studies looking at the associations of polymorphisms in the dopamine transporter (DAT1) with new psychoactive substance use. New psychoactive substances are drugs designed to mimic established drugs, such as cannabis, cocaine, MDMA and LSD [7]. Their conclusion supported a role for certain DAT1 polymorphisms being associated with higher levels of neuroticism and new psychoactive substance use.

This special issue aimed to bring together researchers of substance use disorder across various drugs of abuse and experimental systems to describe the current status of addiction genetics. Psychiatric disorders are known to be complex traits due to their polygenic nature. The articles in this special issue covered various topics and provided diverse insights to direct future SUD genetics research. The growing number of pathways, genes, proteins, and molecules that appear to be involved in SUD is of special interest. We anticipate that this Special Issue will help researchers search for additional genetic associations that will help refine our understanding of the etiology of this complex disease.

Funding: This work was supported by R01DA048890, U01DA043809, P50DA039841 and R01DA059060.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Chunduri, A.; Watson, P.M.; Ashbrook, D.G. New Insights on Gene by Environmental Effects of Drugs of Abuse in Animal Models Using GeneNetwork. *Genes* **2022**, *13*, 614. [CrossRef] [PubMed]
2. Rizk, A.A.; Jenkins, B.W.; Al-Sabagh, Y.; Hamidullah, S.; Reitz, C.J.; Rasouli, M.; Martino, T.A.; Khokhar, J.Y. The Impact of Sex, Circadian Disruption, and the Clock (Delta19/Delta19) Genotype on Alcohol Drinking in Mice. *Genes* **2022**, *13*, 701. [CrossRef] [PubMed]
3. Abdi, H.; Williams, L.J. Principal component analysis. *WIREs Comput. Stat.* **2010**, *2*, 433–459. [CrossRef]
4. Sambo, D.; Goldman, D. Genetic Influences on Fetal Alcohol Spectrum Disorder. *Genes* **2023**, *14*, 195. [CrossRef]
5. Fedorenko, O.Y.; Mikhalitskaya, E.V.; Toshchakova, V.A.; Loonen, A.J.M.; Bokhan, N.A.; Ivanova, S.A. Association of PIP4K2A Polymorphisms with Alcohol Use Disorder. *Genes* **2021**, *12*, 1642. [CrossRef]
6. Chmielowiec, J.; Chmielowiec, K.; Masiak, J.; Pawlowski, T.; Larysz, D.; Grzywacz, A. Analysis of Relationships between DAT1 Polymorphism Variants, Personality Dimensions, and Anxiety in New Psychoactive Substance (Designer Drug) (NPS) Users. *Genes* **2021**, *12*, 1977. [CrossRef]
7. Alcohol and Drug Foundation. New Psychoactive Substances. Available online: https://adf.org.au/drug-facts/new-psychoactive-substances/ (accessed on 9 June 2023).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Lavezzo *et al. BMC Infectious Diseases* 2013, **13**:554
http://www.biomedcentral.com/1471-2334/13/554

BMC Infectious Diseases

RESEARCH ARTICLE **Open Access**

Genomic comparative analysis and gene function prediction in infectious diseases: application to the investigation of a meningitis outbreak

Enrico Lavezzo[1], Stefano Toppo[1], Elisa Franchin[1,2], Barbara Di Camillo[3], Francesca Finotello[3], Marco Falda[1], Riccardo Manganelli[1,2], Giorgio Palù[1,2*] and Luisa Barzon[1,2*]

Abstract

Background: Next generation sequencing (NGS) is being increasingly used for the detection and characterization of pathogens during outbreaks. This technology allows rapid sequencing of pathogen full genomes, useful not only for accurate genotyping and molecular epidemiology, but also for identification of drug resistance and virulence traits.

Methods: In this study, an approach based on whole genome sequencing by NGS, comparative genomics, and gene function prediction was set up and retrospectively applied for the investigation of two *N. meningitidis* serogroup C isolates collected from a cluster of meningococcal disease, characterized by a high fatality rate.

Results: According to conventional molecular typing methods, all the isolates had the same typing results and were classified as outbreak isolates within the same *N. meningitidis* sequence type ST-11, while full genome sequencing demonstrated subtle genetic differences between the isolates. Looking for these specific regions by means of 9 PCR and cycle sequencing assays in other 7 isolates allowed distinguishing outbreak cases from unrelated cases. Comparative genomics and gene function prediction analyses between outbreak isolates and a set of reference *N. meningitidis* genomes led to the identification of differences in gene content that could be relevant for pathogenesis. Most genetic changes occurred in the capsule locus and were consistent with recombination and horizontal acquisition of a set of genes involved in capsule biosynthesis.

Conclusions: This study showed the added value given by whole genome sequencing by NGS over conventional sequence-based typing methods in the investigation of an outbreak. Routine application of this technology in clinical microbiology will significantly improve methods for molecular epidemiology and surveillance of infectious disease and provide a bulk of data useful to improve our understanding of pathogens biology.

Keywords: *Neisseria meningitidis*, Whole genome sequencing, Next generation sequencing, capsule locus, Comparative genomics, 454 pyrosequencing, Meningitis outbreak, Molecular epidemiology, Gene function prediction

Background

The recent success of next generation sequencing technologies (NGS), that are now available in many laboratories and sequencing centers, is deeply impacting our capability to investigate biological samples in many fields of research, including microbiology and virology [1,2]. In this regard, the main applications fostered by NGS are the identification of novel pathogens, the metagenomics analysis of biological and environmental niches, the investigation of transcriptomes, and the characterization of full-length microbial and viral genomes for a wide range of purposes, including phylogenetic and epidemiological studies and the characterization of drug-resistant or immune-escaping mutants [3-6].

For each of these applications, a plethora of bioinformatics tools is available, both academic and commercial, that has to be applied and coordinated depending on the specific needs. In this study we report the application of an analysis pipeline which combines genomic comparative

---

* Correspondence: giorgio.palu@unipd.it; luisa.barzon@unipd.it
[1] Department of Molecular Medicine, University of Padova, Padova, Italy
[2] Regional Reference Laboratory for Infectious Diseases, Microbiology and Virology Unit, Padova University Hospital, Padova, Italy
Full list of author information is available at the end of the article

BioMed Central

© 2013 Lavezzo et al.; licensee BioMed Central Ltd. This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

analysis and gene function prediction to finely characterize microbial genomes. The pipeline was applied to the retrospective analysis of a group of *Neisseria meningitidis* isolates collected during an outbreak. Epidemiological, clinical, and surveillance data on this outbreak, which occurred in north-eastern Italy during the 2007–2008 winter and was characterize by a high fatality rate, have been previously reported [7,8]. The results of this study highlight the essential contribution of whole genome sequencing, performed by NGS technology, to distinguish outbreak cases, i.e., related cases with a common epidemiological source, from clusters of temporary and geographically proximate but unrelated cases. In addition, genomic comparative analysis and gene function prediction led to the identification of genetic changes in the capsule locus that could have contributed to pathogenicity.

Methods

N. meningitidis isolates

N. meningitidis isolates of an outbreak which occurred in Veneto Region (north-eastern Italy) in December 2007-January 2008 were collected by local hospital laboratories and sent to the Regional Reference Laboratory at Padua University Hospital for confirmation, phenotypic characterization, and molecular typing. The outbreak strains analyzed in the present study included *N. meningitidis* isolates from seven patients (mean age 23 year, range 15–33 years) from a relatively small geographical area, who had disease onset between December 13, 2007 and January 4, 2008.

The study was approved by the Ethics Committee of Padova University Hospital (protocol no. 53503).

Phenotypic and genotypic characterization of *N. meningitidis* isolates

Serogrouping, which was performed by slide agglutination using commercial antisera (Remel Europe Ltd, Dartford, UK), classified all the 7 isolates as serogroup C. All isolates were fully susceptible to penicillin, rifampicin, ceftriaxone, and ciprofloxacin. Pulsed-field gel electrophoresis analysis gave the same electrophoresis pattern for all isolates, indicating their relatedness. Molecular characterization by MLST, performed according to Maiden *et al.* [9], demonstrated that all isolates belonged to the same sequence type and clonal complex ST-11 (all isolates had the following MLST alleles: *abcZ* 2, *adk* 3, *aroE* 4, *fumC* 3, *gdh* 8, *pdhC* 4, *pgm* 6). In addition, sequencing of *N. meningitidis PorA* variable regions 1 and 2 [10] confirmed that all isolates had the same *PorA* subtype 5–1, 10–8.

Whole genome sequencing of *N. meningitidis* isolates by 454 pyrosequencing

Whole genome sequencing of two *N. meningitidis* isolates, the index case (named K1207) and the last case (named S0108) of the outbreak, was performed with the objective of confirming their relatedness and to detect genetic differences between the two strains that could have occurred during the short period of the outbreak. The draft genome sequences of these two *N. meningitidis* isolates were reported in a previous announcement [11].

Genomic DNA was purified from meningococcal isolates using a phenol-chloroform-based method.

Sequencing was performed using a Roche 454 Life Sciences Genome Sequencer FLX platform following the manufacturer's instructions (Roche 454 Life Sciences, Branford, CT, USA). For each sample, 2 different libraries were prepared, a shotgun and a 3 kb paired-end, starting from 5 μg of genomic DNA. The shotgun library was prepared as follows: after nebulization, purification and adaptors ligation, DNA fragments were clonally amplified using the Emulsion PCR Kit I (Roche). For the preparation of the paired end library, genomic DNA was fragmented by hydrodynamic shearing, followed by a size selection step, hairpin adaptors ligation and circularization of fragments. From this step the procedure was similar to the shotgun one, consisting in the nebulization of circular molecules, paired end adaptors ligation and amplification of the library. The clonal amplification was carried out using the Emulsion PCR Kit II (Roche). Sequencing was performed on a GS FLX instrument, using the Standard LR70 Sequencing Kit (Roche). Images were processed using the runAnalysisPipe and runAnalysisPipePairedEnd commands provided with the DataProcessing package (Roche). With respect to the previous genome announcement [11], we performed a new *de novo* assembly with the most recent version of the Newbler software (v.2.6), which is more effective and produces a smaller number of contigs with respect to older versions [12].

Gene identification and comparison between the two *N. meningitidis* genomes and comparison with FAM18 reference genome

The assembled sequences of K1207 and S0108 isolates were compared to each other and with the most similar genome among those already available in RefSeq database, i.e., the FAM18 strain genome (NC_008767). Whole genome level comparison was made with MUMmer3.23 package, to investigate possible large recombination events.

To compare the CDS content between K1207 and S0108 genomes an approach to find the reciprocal blast best hits was implemented. This method is widely used to identify the coding sequences that are supposed to be "orthologs". The following steps were performed:

– download of protein sequences of genes annotated on each of the following complete *N. meningitidis* genomes, spanning different serogroups: FAM18 (NC_008767) and 053442 (NC_010120) from serogroup C, Z2491 (NC_003116) from serogroup

A, MC58 (NC_003112) from serogroup B and the capsule null strain alpha14 (NC_013016), for which the serogroup cannot be defined.
- clustering of protein sequences from the previous step at 90% of sequence identity, using cd-hit [13,14], in order to eliminate redundancy and create a reference proteome.
- *de novo* gene prediction on isolates K1207 and S0108 using Glimmer3.02 [15] and extraction of the corresponding protein sequences.
- search of protein sequences belonging to each isolate against this reference proteome using blastp [16]; since all analyzed protein sequences belong to the same species, only matches with at least 90% of sequence identity and covering at least 90% of sequence length were considered. The threshold was chosen after testing different cutoff values: relaxing this parameter did not produce substantial changes in the number of shared CDSs, while tightening the percentage identity cutoff resulted in a rapid decrease in the number of shared CDSs (see Additional file 1: Figure S1). Nonetheless, this high stringency led to the loss of many similar genes whose function is identical.
- starting from blastp output, a table of matches was created in which, for each protein, “0” or “1” indicated its absence or presence in the different genomes. From this table it was possible to extract different lists of genes, common to both genomes, or specific to one or another.

To better characterize the differences between the two isolates, function prediction of unknown genes was done:

- genes with unknown function were analyzed by means of the function prediction tool Argot2 [17-19]. This algorithm was developed by our bioinformatics group and allows to retrieve scored GO annotations for a given gene or CDS, starting from its nucleotide or protein sequence.

The workflow of the analysis pipeline is summarized in Figure 1.

Validation of genetic differences among *N. meningitidis* isolates by target-specific PCR and cycle sequencing

Some of the genes found to be present in only one of the two *N. meningitis* isolates were selected for validation with target-specific PCR and conventional cycle sequencing. A detailed list of investigated genes is reported in Table 1, along with brief gene function descriptions. These genes were also screened in the other five *N. meningitidis* isolates from the outbreak and in sporadic *N. menigitidis* isolates, with identical multilocus sequence type and *porA* genotype, which were collected in Veneto Region in 2007 and 2008.

Results

Pyrosequencing results

The index and the last cases of a *Neisseria meningitidis* outbreak, named K1207 and S0108 respectively, were sequenced in a 454 FLX platform. For K1207 genome, a total of 257,909 shotgun reads were obtained, with average length of 237 bases, and 63,593 paired end reads with an average distance of 3,113 nucleotides. For genome S0108, the shotgun reads obtained were 241,983 with an average length of 235 bases, while paired end reads were 50,658 with an average distance of 2560 bases. The average depth was 27 to 28 fold for both genomes. Raw reads were assembled with Newbler2.6.

Characteristics of sequenced genomes and plasmids of *N. meningitidis* isolates

The genome sequences of the two *N. meningitidis* serogroup C ST-11 isolates were very similar to *N. meningitidis* serogroup C FAM18 strain, which is a representative of the ST-11/ET-37 clonal complex and whose sequence is completely finished and deposited in RefSeq (NC_008767). A deeper analysis revealed that the two isolates belonged to the ET-15 variant of the ST-11 complex, as demonstrated by the presence of the IS*1301* insertion element [20] and the ET-15 fumarate hydratase gene [21].

The sequence of a 7-kb plasmid, identical in both isolates and very similar to pJS-B (RefSeq accession number NC_004758) was also obtained. Draft genome sequences have been deposited in GenBank (accession numbers ADWM02000000 and ADWN02000000 for K1207 and S0108, respectively).

Comparison of *N. meningitidis* serogroup C genomes through analysis of their coding sequence content

To screen for the presence of genetic variations between the two isolates, the K1207 and S0108 genomes were compared by assessing their coding sequence (CDS) content, looking for shared and genome-specific CDSs. The latter were defined as CDSs for which a reciprocal match between the two isolates was not detected. Following the criteria described in the methods section, 2011 CDSs were identified as present in both genomes, 11 CDSs were exclusively present in K1207 genome, while 10 CDSs were present in S0108 genome only. Since most unique CDSs encoded hypothetical proteins of unknown function, we performed a function prediction step by using Argot^2 webservice, an in-house developed tool [17-19]; the complete report of predicted functional annotations is available in Additional file 1: Figure S1. These genes were predicted to encode proteins belonging to the membrane compartment, involved in capsular polysaccharide biosynthetic process, or adhesins and transport proteins.

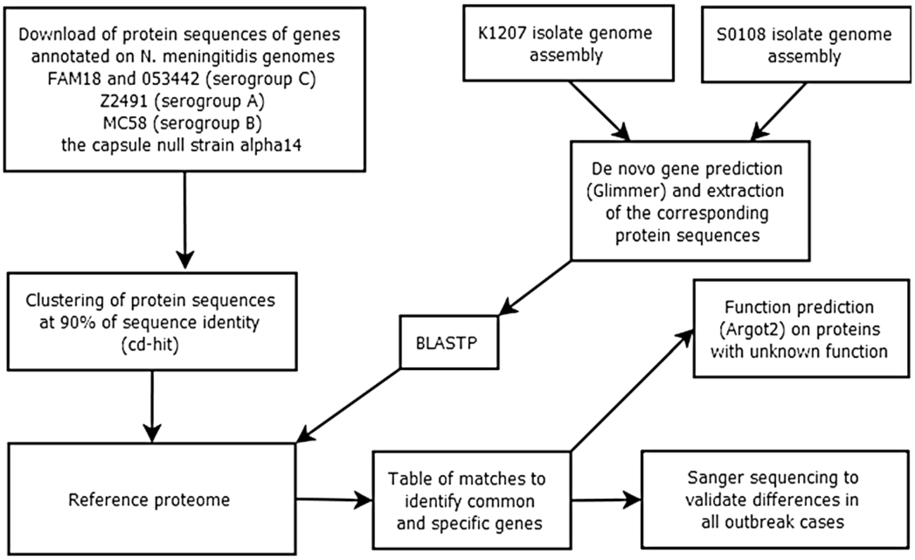

Download of protein sequences of genes annotated on N. meningitidis genomes FAM18 and 053442 (serogroup C) Z2491 (serogroup A) MC58 (serogroup B) the capsule null strain alpha14
K1207 isolate genome assembly
S0108 isolate genome assembly
De novo gene prediction (Glimmer) and extraction of the corresponding protein sequences
Clustering of protein sequences at 90% of sequence identity (cd-hit)
BLASTP
Function prediction (Argot2) on proteins with unknown function
Reference proteome
Table of matches to identify common and specific genes
Sanger sequencing to validate differences in all outbreak cases

Figure 1 Analysis algorithm. Diagram of the analysis workflow applied to the investigation of a meningitis outbreak based on genomic comparative analysis and gene function prediction.

Screening for genetic differences in other *N. meningitidis* serogroup C ST-11 isolates of the outbreak

Since the dissimilarities between the two *N. meningitidis* isolates could be the result of genetic variations of the same strain within the outbreak, we selected a subset of these differences for further investigation and validation by using target-specific PCR and cycle sequencing. This analysis was performed in the two sequenced *N. meningitidis* K1207 and S0108 strains for confirmation and in the other five isolates of the outbreak (i.e., isolates B-F in Table 1), as well as in other two ST-11 isolates (i.e., isolates A and G in Table 1), which were collected from sporadic meningitis cases that occurred in the same area in November 2007 and in April 2008. The results of this analysis confirmed findings provided by NGS and bioinformatics comparison for the K1207 and the S0108 isolates and, unexpectedly, they demonstrated that the S0108 isolate did not belong to the outbreak cluster. In fact, while the other five isolates from the same outbreak shared the same sequences of the outbreak index case K1207, the CDSs of S0108 were found in the strains isolated from sporadic meningitis cases, but differed from those of the outbreak isolates (Table 1). Thus, these results demonstrate that the outbreak was caused by the K1207 strain, while the S0108 strain was circulating in the same region during the same season and caused sporadic meningitis cases.

Analysis of the capsule locus

Comparison among K1207, S0108, and FAM18 *N. meningitidis* genomes at whole genome level and for CDS content demonstrated that most genetic variations occurred in the capsule locus. The results of whole genome comparison between K1207 and S0108 are illustrated in Figure 2, where an inversion in the capsule locus of S0108 is demonstrated. The whole genome alignment is shown with a dot plot matrix, where a dot is drawn each time there is a match between two regions of the genomes. Ideally, a perfect alignment would result in a continuous diagonal line from the start to the end of genomes; the presence of a chromosome inversion is highlighted by a break in the diagonal, with the change in dots direction indicating an inverse complement alignment. A map summarizing genetic changes among K1207, S0108 and FAM18 genomes is shown in Figure 3, where the blue chiasma indicates the same inversion present in S0108 with respect to FAM18.

The K1207 genome had remarkable changes in CDS content and nucleotide sequence of the capsule locus in comparison with both S0108 and FAM18 strains, which were consistent with multiple events of recombination and horizontal gene transfer. Genetic variations in the capsule locus of K1207 affected (i) the *refB* gene, which had a deletion in 3′, (ii) the duplicated *galE-rfbBAC* operon (for lipooligosaccharide biosynthesis) and the downstream *lipAB* operon (for phospholipid modification of the capsular polysaccharide, required for its translocation to the cell surface), which had genes with markedly different nucleotide sequences and the insertion of novel CDSs (i.e., *SacB*, which we predicted to encode a polysaccharide phosphotransferase, and rfbC/D which encodes a TDP-deoxymannose-dehydratase); (iii) and the *oatC* gene (that encodes polysialic acid-specific O-acetyltransferase), which was interrupted by the insertion of a mobile element of the

Table 1 Gene-specific PCR and Sanger sequencing validation of differences detected by NGS

CDS ID	*N. meningitidis* serogroup C isolates								
	A	**K1207**	**B**	**C**	**D**	**E**	**F**	**S0108**	**G**
	Nov 2007	**Dec 2007**	**Dec 2007**	**Dec 2007**	**Dec 2007**	**Dec 2007**	**Dec 2007**	**Jan 2008**	**Apr 2008**
K1207-specific CDSs									
GI:352289428 (SacB, capsular polysaccharide phosphotransferase)	-	+	+	+	+	+	+	-	-
GI:254805876 (putative endonuclease)	-	+	+	+	+	+	+	-	-
GI:388935 (TDP-deoxymannose-dehydratase)	-	+	+	+	+	+	+	-	-
GI:254673887 (lipA5, capsule polysaccharide modification protein)	-	+	+	+	+	+	+	-	-
GI:254671124 (hypothetical protein)	-	+	+	+	+	+	+	-	-
S0108-specific CDSs									
GI:121635673 (putative pilin)	+	-	-	-	-	-	-	+	+
GI:254673874 (hypothetical protein)	+	-	-	-	-	-	-	+	+
GI:254671515 (LipA, Capsule polysaccharide export protein)	+	-	-	-	-	-	-	+	+
GI:385323246 (hypothetical protein)	+	-	-	-	-	-	-	+	+
MLST	ST-11/ET-37	ST-11/ET-37	ST-11/ET-37	ST-11/ET-37	ST-11/ET-37	ST-11/ET-37	ST-11/ET-37	ST-11/ET-37	ST-11/ET-37
PorA VR1/2	P1.5-1, 10-8	P1.5-1,10-8	P1.5-1,10-8	P1.5-1,10-8	P1.5-1,10-8	P1.5-1,10-8	P1.5-1,10-8	P1.5-1,10-8	P1.5-1,10-8

+ present, by PCR; - not present, by PCR.

The target CDSs that were specifically detected in the K1208 and S0108 genomes by 454 pyrosequencing were validated by means of PCR and Sanger sequencing. Then, they were searched also in *N. meningitidis* isolates A-G, which represent all the *N. meningitidis* serogroup C isolates classified as ST-11/ET-15 that were collected in Veneto Region in 2007–2008. Isolates B-F show the same pattern of K1207 and are classified as “outbreak cases”; isolates A and G are sporadic cases and are similar to S0108.

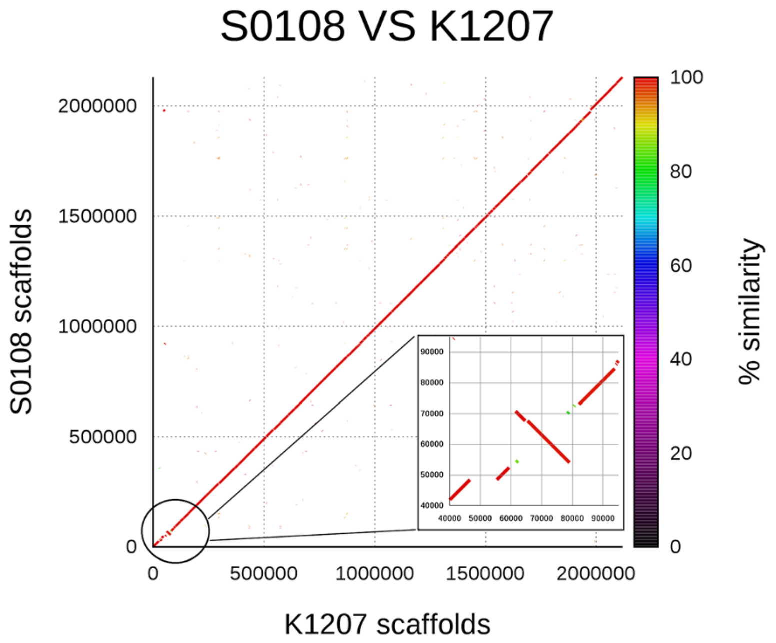


Figure 2 Genome-wide comparison of samples K1207 and S0108. Whole genome alignment between isolate K1207 vs. isolate S0108. The inset highlights the 20 Kb inversion in the capsule locus region between the two isolates. The dot plot was generated with MUMmer3.23 (http://mummer.sourceforge.net) and indicates matching sequences in forward and reverse direction.

IS4 transposase family. Search in the GenBank database demonstrated that the *SacB* gene was present only in the genome of the carriage *N. menigitidis* strain alpha275 (Genbank AM889138), but not in pathogenic strains. In addition, we found this gene also in *Neisseria mucosa*, suggesting that the rearranged region of the capsule locus in K1207 might have been acquired by horizontal gene transfer from nonpathogenic meningococci or other commensal *Neisseria* species.

The CDS content of the S0108 genome showed a high degree of similarity with FAM18 genome, with the exception of a large inversion in the capsule locus of a region of approximately 20 Kb included between the duplicated *rfbC* and *rfbC2* genes (Figures 2 and 3). In addition, the

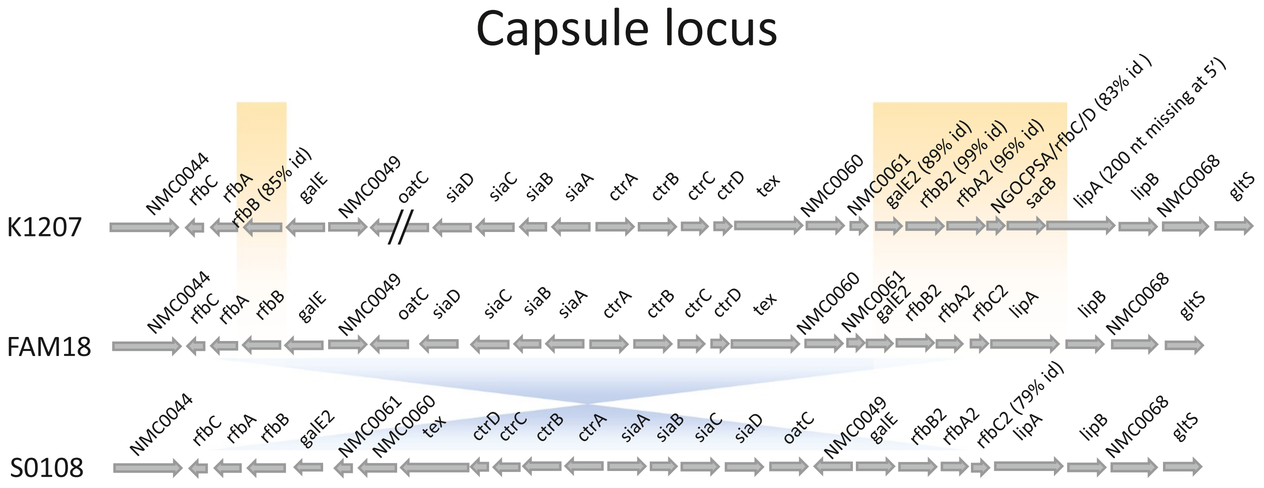

Capsule locus
K1207
NMC0044
rfbC
rfbA
rfbB (85% id)
galE
NMC0049
oatC
siaD
siaC
siaB
siaA
ctrA
ctrB
ctrC
ctrD
tex
NMC0060
NMC0061
galE2 (89% id)
rfbB2 (99% id)
rfbA2 (96% id)
NGOCPSA/rfbC/D (83% id)
sacB
lipA (200 nt missing at 5')
lipB
NMC0068
gltS
FAM18
NMC0044
rfbC
rfbA
rfbB
galE
NMC0049
oatC
siaD
siaC
siaB
siaA
ctrA
ctrB
ctrC
ctrD
tex
NMC0060
NMC0061
galE2
rfbB2
rfbA2
rfbC2
lipA
lipB
NMC0068
gltS
S0108
NMC0044
rfbC
rfbA
rfbB
galE2
NMC0061
NMC0060
tex
ctrD
ctrC
ctrB
ctrA
siaA
siaB
siaC
siaD
oatC
NMC0049
galE
rfbB2
rfbA2
rfbC2 (79% id)
lipA
lipB
NMC0068
gltS

Figure 3 Capsule locus modifications. Map of the capsule locus of K1207 and S0108 *N. meningitidis* isolates in comparison with FAM18 strain. Changes in CDS content between K1207 and FAM18 are highlighted in yellow, with the percentage similarity of amino acid sequences reported in parenthesis; the 20 kb inversion in the S0108 genome is indicated in blue.

nucleotide sequence of the *rfbC2* gene in the FAM18 genome was different from the corresponding gene in K1207 and S0108 genomes.

Discussion

In this study, an approach based on whole genome sequencing by NGS technology, comparative genomics, and gene function prediction was set up and applied to investigate *N. meningitidis* isolates, collected during an outbreak, that appeared identical by standard molecular typing methods based on PFGE, MLST, and *porA* VR1/VR2 sequencing. Whole genome sequencing demonstrated genetic changes between the fully sequenced isolates: then, a specific PCR based strategy applied to other samples from the same outbreak, followed by Sanger sequencing, allowed to distinguish outbreak cases from other temporal and geographically proximate but unrelated cases. This approach was designed in order to be feasible with a small budget. Nonetheless, the continuous decrease of sequencing costs, due to the improvement in the throughput of existing sequencing platforms and the simultaneous development of new technologies, will probably foster a wider employ of next generation sequencing in diagnostics and surveillance activities.

Comparative genomics analysis led to the identification and characterization of genetic differences among strains that might have been relevant for virulence and pathogenicity. In fact, in comparison with the *N. meningitidis* serogroup C ST-11/ET-37 reference strain FAM18 and with the S0108 isolate, the hyper-virulent K1207 outbreak isolate had marked changes in the capsule locus, consistent with recombination and horizontal acquisition of a set of genes involved in capsule biosynthesis from other meningococci. Although experimental verification is required, we hypothesize that these genetic changes might have enhanced transmissibility and invasiveness or even changed antigens involved in host immunity. In this regard, acquisition of novel genotypes at antigen-encoding loci has been reported to be probably the mechanism of emergence of hyper-virulent *N. meningitidis* ST-11 complex strains [22].

Homologous recombination and horizontal gene transfer are quite common in pathogenic *N. meningitidis* strains [23] and frequently affect virulence genes, including the capsule biosynthesis locus which may cause capsule switching [24], genes targeted by vaccines, and antibiotic susceptibility genes [25]. Considering this high variability, the conventional genotyping methods that are used to investigate outbreaks and to monitor the circulation of new variants of hyper-virulent strains may not be accurate enough. Several recent studies have demonstrated the power of whole genome sequencing by NGS technology to improve the identification of the geographical and evolutionary origin of an outbreak and to distinguish cases that belong to the outbreak from cases that do not [26].

The application of NGS technologies for the investigation of outbreaks in real time is becoming feasible thanks to the development of semi-automatic pipelines, as proposed in some recent publications [27-29]. In particular, Vogel *et al.* [27] demonstrated how to apply NGS in clinical practice, focusing on target genes of conventional typing methods (i.e., MLST, *porA* and *fetA* typing) and antimicrobial resistance gene testing (i.e., *penA* and *rpoB*). This approach can be expanded, as shown by Jolley *et al.* [28,29], on different subsets of genes, including the standard MLST loci, the ribosomal genes, personalized lists of specific loci, and whole sets of loci annotated on a complete reference genome. Such tools, which come along web platforms and interactive easy-to-use interfaces, are certainly essential instruments that will allow the introduction of whole genome sequencing in diagnostic routine. Instead, a limitation of their application in research is the capability to identify new determinants, not yet known and annotated [3]. The function prediction step integrated in our approach can provide valuable information to improve our understanding on unknown features, such as horizontal gene transfer and genome rearrangements. Though requiring experimental validation, these predictions could be useful to drive the experimental design and produce annotations: this new information will, on the one hand, enrich public databases and increase our knowledge about bacterial biodiversity and pathogenic mechanisms, and on the other hand it will be available for future routine diagnostic access.

Conclusion

In conclusion, this study represents a practical example on the application of NGS technology to the investigation of an outbreak and how this technology may be helpful in the identification of outbreak cases with a direct clinical impact on the containment measures being taken. This study also shows how the avalanche of genomic information provided by NGS technology and bioinformatics analysis may contribute to our understanding of human pathogens.

Additional file

Additional file 1: Figure S1. GO function predictions for proteins specific to one of the two isolates. Complete list of GO terms predicted with Argot2 webserver [30].

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

EF carried out the wet lab procedures; EL, ST, FF, MF, and BDC performed the bioinformatics analyses; RM, GP and LB conceived the study and helped in its design and coordination; LB, ST and EL wrote the manuscript. All authors read and approved the final manuscript.

Author details
[1]Department of Molecular Medicine, University of Padova, Padova, Italy. [2]Regional Reference Laboratory for Infectious Diseases, Microbiology and Virology Unit, Padova University Hospital, Padova, Italy. [3]Department of Information Engineering, University of Padova, Padova, Italy.

Received: 3 May 2013 Accepted: 13 November 2013
Published: 19 November 2013

References

1. Barzon L, Lavezzo E, Costanzi G, Franchin E, Toppo S, Palù G: **Next-generation sequencing technologies in diagnostic virology.** *J Clin Virol* 2013, **13:**00086–3. doi: S1386-6532.
2. Barzon L, Lavezzo E, Militello V, Toppo S, Palù G: **Applications of next-generation sequencing technologies to diagnostic virology.** *Int J Mol Sci* 2011, **12:**7861–7884.
3. Didelot X, Bowden R, Wilson DJ, Peto TE, Crook DW: **Transforming clinical microbiology with bacterial genome sequencing.** *Nat Rev Genet* 2012, **13:**601–612.
4. Kahvejian A, Quackenbush J, Thompson JF: **What would you do if you could sequence everything?** *Nat Biotechnol* 2008, **26:**1125–1133.
5. Kostic AD, Ojesina AI, Pedamallu CS, Jung J, Verhaak RG, Getz G, Meyerson M: **PathSeq: software to identify or discover microbes by deep sequencing of human tissue.** *Nat Biotechnol* 2011, **29:**393–396.
6. Frasson I, Lavezzo E, Franchin E, Toppo S, Barzon L, Cavallaro A, Richter SN, Palù G: **Antimicrobial treatment and containment measures for an extremely drug-resistant Klebsiella pneumoniae ST101 isolate carrying pKPN101-IT, a novel fully sequenced bla(KPC-2) plasmid.** *J Clin Microbiol* 2012, **50:**3768–3772.
7. Ferro A, Baldo V, Cinquetti S, Corziali P, Gallo G, Lustro G, Paludetti P, Menegon T, Baldovin T, Palù G, Trivello R: **Outbreak of serogroup C meningococcal disease in Veneto region, Italy.** *Euro Surveill* 2008, **13**(2):8008.
8. Fazio C, Neri A, Tonino S, Carannante A, Caporali MG, Salmaso S, Mastrantonio P, Stefanelli P: **Characterisation of Neisseria meningitidis C strains causing two clusters in the north of Italy in 2007 and 2008.** *Euro Surveill* 2009, **14**(16):19179.
9. Maiden MC, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, Zhang Q, Zhou J, Zurth K, Caugant DA, Feavers IM, Achtman M, Spratt BG: **Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms.** *Proc Natl Acad Sci USA* 1998, **95:**3140–3145.
10. Suker J, Feavers IM, Achtman M, Morelli G, Wang JF, Maiden MC: **The porA gene in serogroup A meningococci: evolutionary stability and mechanism of genetic variation.** *Mol Microbiol* 1994, **12:**253–265.
11. Lavezzo E, Toppo S, Barzon L, Cobelli C, Di Camillo B, Finotello F, Franchin E, Peruzzo D, Toffolo GM, Trevisan M, Palù G: **Draft genome sequences of two Neisseria meningitidis serogroup C clinical isolates.** *J Bacteriol* 2010, **192:**5270–5271.
12. Finotello F, Lavezzo E, Fontana P, Peruzzo D, Albiero A, Barzon L, Falda M, Di Camillo B, Toppo S: **Comparative analysis of algorithms for whole-genome assembly of pyrosequencing data.** *Brief Bioinform* 2012, **13:**269–280.
13. Li W, Jaroszewski L, Godzik A: **Clustering of highly homologous sequences to reduce the size of large protein database.** *Bioinformatics* 2001, **17:**282–283.
14. Li W, Jaroszewski L, Godzik A: **Tolerating some redundancy significantly speeds up clustering of large protein databases.** *Bioinformatics* 2002, **18:**77–82.
15. Delcher AL, Bratke KA, Powers EC, Salzberg SL: **Identifying bacterial genes and endosymbiont DNA with Glimmer.** *Bioinformatics* 2007, **23:**673–679.
16. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389–3402.
17. Fontana P, Cestaro A, Velasco R, Formentin E, Toppo S: **Rapid annotation of anonymous sequences from genome projects using semantic similarities and a weighting scheme in gene ontology.** *PLoS One* 2009, **4:**e4619.
18. Falda M, Toppo S, Pescarolo A, Lavezzo E, Di Camillo B, Facchinetti A, Cilia E, Velasco R, Fontana P: **Argot2: a large scale function prediction tool relying on semantic similarity of weighted Gene Ontology terms.** *BMC Bioinformatics* 2012, **13**(Suppl 4):S14.
19. Radivojac P, Clark WT, Oron TR, Schnoes AM, Wittkop T, Sokolov A, Graim K, Funk C, Verspoor K, Ben-Hur A, Pandey G, Yunes JM, Talwalkar AS, Repo S, Souza ML, Piovesan D, Casadio R, Wang Z, Cheng J, Fang H, Gough J, Koskinen P, Törönen P, Nokso-Koivisto J, Holm L, Cozzetto D, Buchan DW, Bryson K, Jones DT, Limaye B, *et al*: **A large-scale evaluation of computational protein function prediction.** *Nat Methods* 2013, **10:**221–227.
20. Elias J, Vogel U: **IS1301 fingerprint analysis of Neisseria meningitidis strains belonging to the ET-15 clone.** *J Clin Microbiol* 2007, **45:**159–167.
21. Vogel U, Claus H, Frosch M, Caugant DA: **Molecular basis for distinction of the ET-15 clone within the ET-37 complex of Neisseria meningitidis.** *J Clin Microbiol* 2000, **38:**941.
22. Jolley KA, Wilson DJ, Kriz P, McVean G, Maiden MC: **The influence of mutation, recombination, population history, and selection on patterns of genetic diversity in Neisseria meningitidis.** *Mol Biol Evol* 2005, **22:**562–569.
23. Schoen C, Tettelin H, Parkhill J, Frosch M: **Genome flexibility in Neisseria meningitidis.** *Vaccine* 2009, **27**(Suppl 2):B103–B111.
24. Swartley JS, Marfin AA, Edupugnanti S, Liu LJ, Cieslak P, Perkins B, Wenger JD, Stephens DS: **Capsule switching of Neisseria meningitidis.** *Proc Natl Acad Sci USA* 1997, **94:**271–276.
25. Hao W, Ma JH, Warren K, Tsang RS, Low DE, Jamieson FB, Alexander DC: **Extensive genomic variation within clonal complexes of Neisseria meningitidis.** *Genome Biol Evol* 2011, **3:**1406–1418.
26. Didelot X: **Genomic analysis to improve the management of outbreaks of bacterial infections.** *Expert Rev Anti Infect Ther* 2013, **11:**335–337.
27. Vogel U, Szczepanowski R, Claus H, Jünemann S, Prior K, Harmsen D: **Ion Torrent Personal Genome Machine sequencing for genomic typing of Neisseria meningitidis for rapid determination of multiple layers of typing information.** *J Clin Microbiol* 2012, **50:**1889–1894.
28. Jolley KA, Maiden MC: **Automated extraction of typing information for bacterial pathogens from whole genome sequence data: Neisseria meningitidis as an exemplar.** *Euro Surveill* 2013, **18:**20379.
29. Jolley KA, Hill DM, Bratcher HB, Harrison OB, Feavers IM, Parkhill J, Maiden MC: **Resolution of a meningococcal disease outbreak from whole-genome sequence data with rapid Web-based analysis methods.** *J Clin Microbiol* 2012, **50:**3046–3453.
30. *A.r.g.o.t.2, Functional annotation of proteins using the semantic similarity in the Gene Ontology.* http://www.medcomp.medicina.unipd.it/Argot2.

doi:10.1186/1471-2334-13-554
Cite this article as: Lavezzo *et al.*: Genomic comparative analysis and gene function prediction in infectious diseases: application to the investigation of a meningitis outbreak. *BMC Infectious Diseases* 2013 **13**:554.

Submit your next manuscript to BioMed Central and take full advantage of:

- **Convenient online submission**
- **Thorough peer review**
- **No space constraints or color figure charges**
- **Immediate publication on acceptance**
- **Inclusion in PubMed, CAS, Scopus and Google Scholar**
- **Research which is freely available for redistribution**

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central

Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

HHS Public Access
Author manuscript
Nat Rev Nephrol. Author manuscript; available in PMC 2019 February 01.

Published in final edited form as:
Nat Rev Nephrol. 2018 February ; 14(2): 83–104. doi:10.1038/nrneph.2017.167.

Genomic medicine for kidney disease

Emily E. Groopman, **Hila Milo Rasouly**, and **Ali G. Gharavi**
Division of Nephrology, Columbia University College of Physicians and Surgeons, 1150 Saint Nicholas Avenue, Russ Berrie Pavilion #412C, New York, New York 10032, USA.

Abstract

Technologies such as next-generation sequencing and chromosomal microarray have advanced the understanding of the molecular pathogenesis of a variety of renal disorders. Genetic findings are increasingly used to inform the clinical management of many nephropathies, enabling targeted disease surveillance, choice of therapy, and family counselling. Genetic analysis has excellent diagnostic utility in paediatric nephrology, as illustrated by sequencing studies of patients with congenital anomalies of the kidney and urinary tract and steroid-resistant nephrotic syndrome. Although additional investigation is needed, pilot studies suggest that genetic testing can also provide similar diagnostic insight among adult patients. Reaching a genetic diagnosis first involves choosing the appropriate testing modality, as guided by the clinical presentation of the patient and the number of potential genes associated with the suspected nephropathy. Genome-wide sequencing increases diagnostic sensitivity relative to targeted panels, but holds the challenges of identifying causal variants in the vast amount of data generated and interpreting secondary findings. In order to realize the promise of genomic medicine for kidney disease, many technical, logistical, and ethical questions that accompany the implementation of genetic testing in nephrology must be addressed. The creation of evidence-based guidelines for the utilization and implementation of genetic testing in nephrology will help to translate genetic knowledge into improved clinical outcomes for patients with kidney disease.

Although individual inherited kidney diseases are rare, together they account for approximately 10% of adult end-stage renal disease (ESRD)[1–3] and at least 70% of paediatric[4,5] nephropathy. In addition to known hereditary aetiologies, compelling evidence exists for a genetic contribution across different forms of kidney disease. The `heritability` of glomerular filtration rate is estimated to be ~30–60% in the general population[6,7], and other parameters such as tubular transport of electrolytes similarly show substantial hereditability[8–10]. Moreover, 10–29% of adult patients with ESRD report a positive family history across different ethnicities and aetiologies[11– 13]. Prognosis, disease course, and appropriate management can differ markedly between hereditary and acquired forms of kidney disease, but these forms can be indistinguishable when traditional

---

Correspondence to A.G.G. ag2239@cumc.columbia.edu.

Author contributions

All authors contributed substantially to the research and writing process, including discussing the content of the article and reviewing and editing the manuscript before submission.

Competing interests statement

The authors declare no competing interests.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

diagnostics alone are used[14,15]. Moreover, as early-stage disease is often clinically silent, patients might not present until they reach ESRD. Owing to such late referral and the limitations of noninvasive diagnostic tools, the cause of kidney failure is classified as unknown in more than one in ten patients with ESRD[16–18]. Thus, earlier and more specific diagnoses are needed to enable delivery of effective care[19,20].

Technologies such as `chromosomal microarray` (CMA) and `next-generation sequencing` (NGS) were first introduced as research tools in the past decade and have empowered investigation of genetic contributions to disease through assessing variation across the `genome`[21–23]. These technologies are being increasingly deployed in the clinic, reflecting a shift in the work-up of suspected hereditary disorders in clinical medicine. Traditionally, diagnosis of individuals suspected to have an inherited form of nephropathy involved multiple clinical visits and complex and/or invasive studies (such as renal biopsy or biochemical testing) to identify the most likely aetiology, which the clinician would use to select one or a few associated genes for `genetic testing`. Now, a shift towards `genomic medicine` is occurring, wherein genome-wide testing approaches are used to identify the causal aetiology of disease in a patient, and the genetic variant or variants that are identified are used to guide their clinical care[24]. As sequencing costs are rapidly declining, genome-wide testing has been advocated as a sensitive and ultimately cost-effective first-line diagnostic test[25–27]. Use of such testing could help to surmount the diagnostic challenges posed by the substantial genetic and phenotypic heterogeneity[28] of many hereditary nephropathies.

However, in order to realize the promise of genomic medicine in nephrology, many complex questions must be answered, including how to identify patients for whom genetic testing is indicated, choose the appropriate test, identify causal variants, and translate genetic findings into personalized care. Addressing these questions requires thorough investigation in large patient cohorts of diverse ages, ethnicities, and disease aetiologies. The ethical, legal, and social implications (ELSIs) that accompany genetic testing on an increasingly broad scale must also be a priority.

In this Review, we examine the opportunities and challenges that accompany the shift to genomic medicine for the diagnosis of genetic diseases of the kidney and urinary tract. We discuss the tests that are available for the work-up of suspected hereditary nephropathy and their diagnostic advantages and limitations, the patient populations for whom genetic testing is indicated, the current guidelines for interpretation of sequence data to achieve a molecular diagnosis, and the value of a `genetic diagnosis` in nephrology. We highlight the key issues that must be addressed to enable implementation of genetic findings in nephrology clinical practice and discuss how they might be surmounted.

Diagnostic genetic testing

Diagnostic genetic testing aims to identify the mutation that is causal for disease in an individual patient, but the wealth of variation within the human genome makes this task difficult. The typical human genome contains approximately 3 billion DNA nucleotides, of which ~20 million may be altered without major consequences for an individual's health[29],

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

and ~20,000 genes, of which nearly 4,000 have been implicated in human disease[30]. Sequence changes might occur at any of the nucleotide sites and include `single-nucleotide variants` (SNVs), small `insertions or deletions` involving <5–10 bp (indels), and `structural variants`. Given the abundance of rare, predicted damaging variants in a typical human genome, the risk of falsely attributing causality is high[31]. Thus, a major challenge in genetic diagnostics is to identify which variants are disease-causing mutations. Common modalities for diagnostic genetic testing include `Sanger sequencing`, CMA, and NGS approaches, including `targeted next-generation sequencing panels`, `whole-exome sequencing` (WES) and `whole-genome sequencing` (WGS) (TABLE 1).

Sanger sequencing

Sanger sequencing has high analytical validity in detecting causal SNVs and small (<5–10 bp) insertions or deletions. Thus, this modality remains the gold standard for molecular diagnosis when a single-gene disorder is suspected and for confirmation of NGS findings[32,33]. Sanger sequencing also has utility for genomic regions refractory to NGS, such as those that are highly repetitive, homologous, or guanine–cytosine (GC)-rich. However, as Sanger sequencing is limited to single DNA fragments of <1,000 bp (REF. 34), this modality cannot detect larger structural variants and becomes increasingly costly and time-inefficient as the number of candidate genes increases, limiting its utility for genetically heterogeneous conditions[32,35].

Chromosomal microarray

Historically, testing for genomic disorders — that is, genetic diseases caused by structural variants[36] — involved `karyotyping`, which can detect chromosomal disorders, translocations, and other large genomic imbalances. However, many genomic disorders are caused by `copy number variants` (CNVs) that fall below the 1–2 Mb resolution of karyotyping[37]. CMA enables detection of both small and large CNVs[38,39].

Two major types of CMA are used in the clinical setting: `array comparative genomic hybridization` and `single-nucleotide polymorphism arrays`. Both of these techniques offer excellent genome-wide coverage and use enrichment of probes in clinically relevant regions to enable resolution at the single-exon level[23,40,41]. Owing to this high resolution, CMAs have as much as 10-fold increased diagnostic yield versus karyotyping for intellectual disability, autism, and multiple congenital anomalies[23,42,43] and are now recommended as a first-line diagnostic for these indications[23,44]. Unlike karyotyping, however, CMAs cannot detect `balanced chromosomal rearrangements` or low-grade mosaicism and have limited sensitivity to detect changes in certain regions, such as pseudogenes and repetitive elements. In addition, CMAs can generally resolve the boundaries of a CNV to only ~1–2 kb, hindering accurate determination of its size and of the genes affected[45,46], which are key criteria in diagnostic CNV interpretation[21,45]. Thus, findings need to be interpreted in the context of the coverage of the specific CMA platform utilized and the strength of clinical suspicion for a genomic disorder[44].

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Applications

Congenital anomalies of the kidney and urinary tract (CAKUT) are the leading cause of paediatric nephropathy[5,47] and can present in isolation or as part of a multiorgan syndrome[48]. CMA has been shown to be an effective first-line diagnostic tool among patients with syndromic CAKUT[49,50] and may also have utility for those with nonsyndromic forms. For example, a study of 522 children with renal hypodysplasia identified CNVs that were pathogenic for a known genomic disorder in 55 of 380 (14.5%) patients with isolated genitourinary anomalies[50]. Among all studies published to date, pathogenic CNVs were identified in ~4–10% of patients with CAKUT and appear to be enriched among patients with renal parenchymal defects, explaining up to 10% of these cases[50,51–56]. These studies identified more than 40 distinct genomic disorders in children with kidney disease, with recurrent syndromes including deletions at the 17q12 (Online Mendelian Inheritance in Man (OMIM) 614527), 22q11.2 (OMIM 188400), and 16p11.2 (OMIM 611913) loci. Moreover, CMA analysis of a cohort of 419 children with all-cause chronic kidney disease (CKD) detected diagnostic CNVs in 7.4% of patients[55], a diagnostic yield comparable to that noted for the established indications of developmental delay or prenatal testing[23,43,51]. Importantly, the majority of patients studied were referred for routine evaluation to paediatric nephrology or urology clinics and had not been previously diagnosed with a syndromic form of disease, highlighting the difficulty of detecting genetic syndromes by use of traditional clinical methods.

In many cases, a diagnosis of a genomic disorder reclassifies a patient's disease, with important implications for subsequent clinical management. Moreover, these disorders often have pleiotropic effects, including metabolic, skeletal, and neurological complications, which may initiate unnecessary and prolonged work-up in the absence of a unifying aetiology. Alternately, such manifestations may be erroneously attributed as being secondary to renal dysfunction, leading to unachievable expectations for remission after appropriate nephrological treatment. For example, children with nephropathy are at increased risk of adverse neurocognitive outcomes, but this risk has been attributed to the medical and psychosocial burden of kidney disease[57–59]. However, a study of children with CKD demonstrated that those with a genomic disorder had poorer neurocognitive performance, independent of the severity of kidney disease[60]. These findings suggest that in some patients, neurocognitive deficits result from a genomic disorder that impairs both renal and neuropsychiatric function. Thus, CMA has the potential to explain seemingly disparate clinical features, to frame therapeutic expectations, and to guide treatment approaches. Although further investigation is needed to comprehensively ascertain the prevalence of genomic disorders and indications for CMA among adult patients with kidney disease, the above findings in children strongly support its utility as a first-line tool in the diagnostic evaluation of paediatric nephropathy.

Next-generation sequencing

NGS utilizes targeted capture and massively parallel sequencing to simultaneously assess variation in selected regions of the genome, enabling rapid and cost-effective large-scale genetic investigation[32,34,61,62]. Selected regions can be multiple genes of interest

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

(investigated using targeted NGS panels), all protein-coding regions (investigated using WES), or both coding and non-coding regions (investigated using WGS). Each NGS approach has merits and drawbacks in the current landscape of clinical testing, and selection of the optimal modality is rapidly evolving, reflecting technical capacity, cost-effectiveness, and knowledge of the individual patient and disease. As the technical expenses of sequencing have steadily decreased, the substantial time and monetary costs required for diagnostic interpretation have become the major barriers to systematic implementation in clinical practice[32,63].

Targeted panels

NGS gene panels use targeted enrichment of selected genes to provide rapid and inexpensive sequencing at higher coverage than that achieved with WES or WGS[33,35]. Such panels have been advocated as a first-line test for the molecular diagnosis of inherited nephropathies[4]. In this approach, patients are tested for a set of genes that are commonly associated with the phenotype under consideration; for example, a patient with nephrotic syndrome would be tested using a panel containing genes that are commonly implicated in hereditary forms of this disorder[4].

As NGS panels are quickly becoming a first-line diagnostic test, it is critical that they can accurately detect whether a particular genetic variant is present in the region of interest[64,65]. Organizations such as the American College of Medical Genetics and Genomics (ACMG)[66], British Association for Clinical Genetic Science (ACGS)[67], and European Society of Human Genetics (ESHG)[68] have published technical guidelines for clinical NGS regarding `sequencing coverage and depth` as well as other quality metrics. As certain regions, such as those with high GC content (for example, the first exon of *COL4A3*, which is associated with Alport syndrome (OMIM 104200; 203780) and thin basement membrane disease (OMIM 141200)) and those with high sequence homology (for example, the *PKD1* gene, which is associated with autosomal dominant polycystic kidney disease (OMIM 173900)), are poorly covered by NGS alone, laboratories often incorporate other methods such as Sanger sequencing and long-range PCR to ensure that all targeted regions are comprehensively covered at sufficient depth[33].

As sequencing is selective, targeted panels will not yield incidental findings in genes unrelated to the primary indication for testing, reducing the potential burden of `secondary findings` that would initiate additional clinical testing for patients and physicians. If the targeted panel testing is negative, the clinician can select another panel with broader content or proceed directly to WES or WGS. This sequential procedure may be the most comprehensive and cost-effective approach at present, particularly among patients whose presentation is strongly suggestive of a specific category of genetic disease[33,35,69].

Applications—Targeted gene panels are a sensitive and cost-effective diagnostic for a wide range of kidney disorders, including nephrotic syndrome[70,71], nephrolithiasis[72,73], nephronophthisis-related ciliopathies (NPHP-RC)[74,75], and CAKUT[76,77], albeit with variable disease-specific yield in the context of testing familial and/or paediatric cases. Such targeted testing is particularly well suited to diseases that have fairly low genetic

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

heterogeneity. For example, mutations in three genes, *COL4A3, COL4A4*, and *COL4A5*, cause Alport syndrome (OMIM 104200; 203780; 301050) and the related milder form, thin basement membrane disease (OMIM 141200)[78,79]. Targeted NGS sequencing of these genes detected causal variants in 84 (83%) of 101 patients with a clinical diagnosis of familial haematuria[80]. Supplementation with `multiplex ligation-dependent probe amplification`, CMA, and Sanger sequencing enabled identification of large genetic rearrangements and causal variants in regions that were not well captured by NGS alone.

However, many kidney diseases are genetically heterogeneous, and many nephropathy-associated genes can show clinically disparate presentations, reflecting both genetic and environmental modifiers[28]. In such cases, designing a gene panel that appropriately balances sensitivity and specificity is challenging. Restricting the number of genes in the panel reduces the cost and time needed for testing; however, the panel might require frequent updates as new genes are discovered and previously implicated genes are shown to have weaker disease associations than first reported. Assessing a greater number of genes increases diagnostic sensitivity but can also increase the detection rate of `variants of uncertain significance` (VUS), complicating interpretation and clinical follow-up. For example, targeted sequencing of 23 known genes associated with autosomal forms of CAKUT had an ~8% diagnostic yield: 6% in 17 genes for autosomal dominant forms[76] and 2.5% in 6 genes for autosomal recessive forms[77]. An expanded sequencing panel of 208 genes associated with syndromic or isolated CAKUT, including genes implicated by functional data, identified candidate variants in 151 of 453 (33%) patients as well as 32 VUS in 69 (15%) patients[81]. Further expansion to a 330-gene panel screen detected candidate variants in 122 of 204 (60%) patients with CAKUT but also identified 120 VUS in 89 patients (44%)[82].

For patients with more ambiguous presentations, ‘Mendeliome’ panels, which target all known disease-associated genes, have been suggested as a time and cost-effective first-line test, and the studies to date report high diagnostic yield across a range of clinical indications[83–85]. Such panels also enable detection of `phenotypic expansions` of known genetic diseases but require periodic updates to include newly discovered disease-associated genes. Thus, the diagnostic utility and cost-effectiveness of Mendeliome panels relative to WES and WGS merits further in-depth study.

Whole-exome and whole-genome sequencing

WES and WGS provide more comprehensive testing than targeted NGS panels because they assess variation across the genome. These unbiased approaches have many advantages, including increased sensitivity for diagnosis of disorders with high genetic and/or phenotypic heterogeneity and the ability to achieve a specific diagnosis when traditional clinical methods are unsuccessful. WES and WGS also enable reanalysis of sequence data, which may include recalling variants from raw data, reannotating called variants by use of novel bioinformatics tools, and/or re-examining annotated variants in light of newly discovered gene–disease associations. Various analytical frameworks can also be used to identify novel candidate genes for follow-up study[31,86,87], and such re-examination can lead to additional diagnoses, increasing overall diagnostic yield[88–91].

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Per-base coverage is generally lower with WES and WGS than with targeted panels. However, an *in silico* analysis showed that WES with standard coverage (≥10 times coverage of 90% of bases) was adequate to identify 98.6% of sites previously found to have pathogenic variants by targeted panel[92], suggesting that the sensitivity of WES is sufficient for diagnostic sequencing in most cases. Nevertheless, clinically relevant segments of the genome can be missed when using WES alone[93]. For example, the sites corresponding to ~50% of reported pathogenic variants in the *WT1* gene, which is associated with hereditary nephrotic syndrome (OMIM 256370) and Deny–Drash syndrome (OMIM 194080), were poorly covered across three leading WES capture kits[94]. Further development of sequencing technology will help to increase technical accuracy, achieve more uniform coverage, and decrease costs[61,62]. Thus, increasingly comprehensive genomic sequencing has been predicted to define the future of clinical genetic testing, with targeted panels superseded by WES, which in turn will be overtaken by WGS as a first-line diagnostic[32,35,63].

Whether WES or WGS will prove to be a superior clinical diagnostic tool in the near future is a topic of ongoing debate. As known causal variants for Mendelian disorders overwhelmingly lie in coding regions[95,96], WES has been suggested as a time-efficient and cost-efficient means for clinical diagnosis and genetic discovery[86,97,98] and has been successfully used for a variety of clinical indications. To date, the majority of diagnostic variants identified in clinical WGS investigations have been found in exonic regions[99–102]. Non-coding variants have, however, been implicated in various kidney disorders[103–107]. For example, WGS detected a deep intronic mutation in *DGKE* in two unrelated families with infantile-onset atypical haemolytic uraemic syndrome (OMIM 615008) who had been left undiagnosed by use of WES[103]. Subsequent analysis of patient RNA showed that the variant created a novel splice site that abrogated normal protein function. Intronic mutations resulting in altered splicing have likewise been noted in genetically unresolved cases of Alport syndrome[104], Schimke immune-osseous dysplasia[105] (OMIM 242900), and Gitelman syndrome[106] (OMIM 263800). Sequencing the whole genome also avoids capture bias and provides more complete per-base coverage of coding and non-coding regions[108,109], facilitating accurate detection of variants in genes with highly homologous regions, such as *PKD1* (REF. 110), and of structural variants, such as those found in several patients with Joubert syndrome[111]. Although further study is needed, WGS has been reported to detect causal variants in ~20–40% of patients left undiagnosed by WES and/or CMA[112–114].

Importantly, some types of variants remain wholly refractory to detection using current NGS technologies. For example, causal variants in the *MUC1* gene, which is associated with autosomal dominant tubulointerstitial kidney disease (ADTKD-*MUC1*; OMIM 174000) and contains a highly repetitive, GC-rich sequence, were missed by NGS-based regional capture, WES, and WGS and were identified only by long-range PCR and molecular cloning[115]. A novel assay based on mass spectrometry has been recently developed for diagnosis of ADTKD-*MUC1* (REF. 116), and with the advent of long-read sequencing, NGS-based detection of such regions may become feasible[61].

Given such increased diagnostic and analytical sensitivity, WGS has the potential to provide a single-test solution. However, with higher-performance WES capture platforms and improvements in NGS technology, the extent to which the benefits of the expanded search

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

space outweigh the burdens of sequencing costs, data storage, and clinical interpretation and follow-up remains unclear[117,118].

To date, investigations of the diagnostic yield of genome-wide testing in nephrology have overwhelmingly used WES. To our knowledge, the efficacy of WGS in this field remains to be comprehensively assessed. Thus, we focus below on selected studies demonstrating the utility of WES as a diagnostic tool for various forms of nephropathy.

Testing in an expanding genetic spectrum—WES has been successfully employed for a variety of conditions with high genetic heterogeneity, such as NPHP-RC and nephrotic syndrome. NPHP-RC has >90 known causal genes[119], and the advent of NGS has accelerated the discovery of additional causal genes[120]. Expansion of the genetic search space thus enables identification of mutations in noncanonical genes, increasing diagnostic yield relative to targeted panel testing. Detection rates in testing for NPHP-RC were 12% with a 13-gene panel[121], 21% with a 34-gene panel[74], and ~60–70% with WES[122,123]. Similarly, WES identified causal variants in 49 of 187 (26.2%) paediatric patients with steroid-resistant nephrotic syndrome (SRNS)[124]. Although 30 (61.2%) of the 49 patients with resolved cases had mutations in canonical genes for early-onset SRNS (*NPHS1, NPHS2*, or *WT1*), three had diagnostic variants in genes classically associated with other renal disorders, including *DGKE* (membranoproliferative glomerulonephritis and/or atypical haemolytic uraemic syndrome; OMIM 615008), *COL4A3* (Alport syndrome), and *OCRL* (Dent disease 2; OMIM 300555). These three patients all presented with primary SRNS and focal segmental glomerulosclerosis (FSGS) on renal biopsy, suggesting that the findings did not result from initial clinical misphenotyping.

Moreover, use of NGS has led to the detection of novel genes in many disorders that were previously thought to be highly genetically homogenous. For example, polycystic kidney disease (PKD) was long thought to result from mutations in only three genes, with those in *PKD1* and *PKD2* accounting for the autosomal dominant form (ADPKD; OMIM 173900, 613095) and those in *PKHD1* leading to autosomal recessive disease (ARPKD; OMIM 263200). However, ~7–10% of families with ADPKD lack mutations in *PKD1* or *PKD2* (REFS 125,126), and *PKHD1* mutations are not detected in at least 13% of patients with ARPKD[127,128]. WES of mutation-negative families implicated *GANAB* in ADPKD[129] and *DZIP1L* in ARPKD[130], broadening the genetic spectrum of PKD.

Expanding the phenotypic spectrum—Conversely, WES has demonstrated that many genetic disorders can produce a wider range of phenotypes than previously thought. These phenotypic expansions dispel the classical view that a one-to-one relationship exists between a gene and a disease and challenge traditional clinical classifications[28]. For example, mutations in *COL4A3, COL4A4*, and *COL4A5*, which are associated with Alport syndrome, have been detected among patients with a clinical diagnosis of nephrotic syndrome, expanding the range of phenotypes associated with *COL4A*-mediated nephropathy[131,132]. The variability of phenotypes among patients with mutations in *HNF1B* similarly exemplifies this point. Although *HNF1B* mutations are classically associated with renal cysts and diabetes syndrome (OMIM 137920), patients with these mutations can be nondiabetic or have other, noncystic forms of renal disease[133,134]. For example, patients

Author Manuscript

with *HNF1B*-mediated disease can present with hyperuricaemia and glomerulocystic kidney disease, such that they may be mistakenly diagnosed with ADTKD; others may present with hypomagnesaemia and hypocalciuria, causing an assumption that they have Gitelman syndrome[133]. Moreover, many of the extrarenal features of *HNF1B*-mediated disease, such as hyperuricaemia and hyperparathyroidism, can also occur as secondary complications of renal dysfunction; thus, patients might not be suspected to have a genetic form of nephropathy, especially those who are older and/or have no family history[133].

Other examples include the expansion of variants in *PAX2*, which are classically associated with CAKUT (OMIM 120330), to include hereditary FSGS (OMIM 616002)[135] and of biallelic mutations in *TTC21B* to span both cystic and glomerular disease[136,137]. Similarly, the phenotypic spectrum of *UMOD*-associated kidney disease (OMIM 609886; 162000; 603860) encompasses both tubulointerstitial nephritis and glomerulocystic disease[138], and the presentation of patients with *CLCN5* mutations, causal for Dent disease 1 (OMIM 300009), may range from tubulointerstitial disease with electrolyte imbalances[139,140] to nephrotic-range proteinuria and glomerulosclerosis on renal biopsy[141,142].

Proteinuria and glomerulosclerosis have also been reported among patients with variants in *SLC12A3*, which are associated with Gitelman syndrome, a salt-wasting distal tubulopathy (OMIM 263800)[143–145]. Similarly, mutations in *PARN* cause a telomere syndrome traditionally associated with pulmonary fibrosis and bone marrow failure (OMIM 616371), but a WES study suggests that some patients with these mutations initially present with renal tubulointerstitial fibrosis, expanding the phenotypic spectrum of *PARN*-mediated disease[146].

Resolving undiagnosed disorders—WES has transformed paediatric and neonatal care by providing a means to rapidly resolve undiagnosed cases, informing prognosis and clinical management[22,101,147]. Similarly, WES may have considerable diagnostic value for patients who present with nonspecific renal phenotypes or kidney disease of unknown aetiology (REFS 148–151). For example, one of the first clinical applications of WES was for a neonate who presented with hypokalaemic metabolic alkalosis and was thus suspected to have Bartter syndrome[148]. WES identified no candidate variants at Bartter-associated loci but rather a homozygous substitution mutation at a highly conserved residue in *SLC26A3*, and clinical follow-up confirmed the genetic diagnosis of congenital chloride diarrhoea (OMIM 214700). Putatively pathogenic variants in *SLC26A3* were found in an additional 5 of 39 (13%) patients with presumed Bartter syndrome, supporting the utility of exome-wide analysis in resolving such clinically inscrutable cases.

Subsequent studies have further highlighted this utility. For example, exome analysis of 79 children who presented with increased renal echogenicity on ultrasonography identified causal variants in 63% of these patients[122]. Notably, 36% of the patients with a genetic diagnosis were found to have a disorder other than NPHP-RC, which was the diagnosis suspected on the basis of the ultrasonography results. The researchers hypothesized that the misclassification of patients reflects the nonspecificity of increased renal echogenicity for diagnosis of NPHP-RC. Similarly, WES of 33 consanguineous families diagnosed with CAKUT detected pathogenic variants in 9 (27%) families; notably, 4 (44%) of these families had mutations in genes unassociated with CAKUT, and on clinical follow-up, they were

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

found to have phenotypes concordant with these alternate genetic aetiologies of disease[152]. In many patients, these unexpected diagnoses had important implications for management and therapy, including tight regulation of salt and water intake for nephrogenic diabetes insipidus type 2 (OMIM 125800)[153], cysteamine supplementation for nephropathic cystinosis (OMIM 219800)[154], and combined liver and renal transplantation for primary hyperoxaluria type 1 (OMIM 259900)[155].

Moreover, a pilot study of WES in adults with familial or undiagnosed nephropathy identified diagnostic findings in 22 of 92 (24%) patients[146]. Diagnostic yield was notably high among those with kidney disease of unknown aetiology, with 9 of 16 (56%) patients being diagnosed. These diagnoses encompassed a variety of genetic nephropathies, including autosomal and X-linked forms of Alport syndrome, Dent disease, CHARGE syndrome (OMIM 214800), and *HNF1B*-associated disease. In addition to ending the ‘diagnostic odyssey’, the genetic diagnoses resulting from WES impacted clinical care in many patients, including guiding choice of therapy (for example, steroid avoidance in patients with glomerulonephritis and *COL4A3–5* mutations), advising subsequent work-up and surveillance of associated extrarenal comorbidities (for example, screening for diabetes and liver function in a patient with an *HNF1B* mutation), and informing transplant prognosis and choice of donor (for example, a low risk of disease recurrence and genetic screening for candidate living related donors in a patient with Dent disease). These findings support the diagnostic utility of WES for patients with kidney disease of unknown aetiology, particularly those with familial or early-onset disease, and highlight the need for further research in this field in larger cohorts.

Indications for genetic work-up

With the widening availability and declining costs of sequencing technologies, nephrologists will increasingly incorporate clinical genetic testing into their diagnostic armamentarium. There is a risk that these new technologies will be adopted prematurely, before systematic evidence of their utility has been generated. Hence, there is a need for large, multicentre studies of diverse cohorts to develop evidence-based guidelines regarding the indications and utility of genetic testing in nephrology.

Currently, genetic testing is recommended as part of the diagnostic work-up for patients with paediatric kidney disease, especially among those with nondiagnostic presentations[156]. For adult patients, genetic testing is suggested only for those who are strongly suspected to have a known hereditary form of nephropathy. Other indications for genetic work-up include phenotypes that have a strong hereditary basis, such as CAKUT, for which CMA seems to be a valuable first-line diagnostic modality in addition to NGS-based approaches. Certain clinical situations might also merit genetic testing, such as those in which diagnostic findings would enable patients to avoid undergoing unnecessary invasive procedures (for example, renal biopsy in patients with nephronophthisis) or prevent them from receiving ineffective and costly treatment with substantial adverse effects (for example, steroid therapy in patients with hereditary aetiologies of SRNS). Genetic testing is also advised for females with clinical features and/or a history suggestive of a monogenic X-linked nephropathy, such as X-linked Alport syndrome (OMIM 301050) or Fabry disease (OMIM 301500), because

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

although female carriers of these diseases generally display a milder (often subclinical) phenotype than is seen in males, they can develop severe disease[157–159]. At present, clinicians are generally advised to start with a disease-specific genetic panel and if the results are negative, to proceed to a Mendeliome panel, WES, or WGS[35,66,68].

Genetic testing has also been recommended in the evaluation of potential living kidney donors, with donation contraindicated among those found to have autosomal dominant forms of inherited kidney disease such as ADPKD (OMIM 173900, 613095, 600666) or to share genetic susceptibility factors for atypical haemolytic uraemic syndrome[160]. In such cases, positive findings not only guide choice of donor but also might inform renal prognosis; for example, among patients with ADPKD owing to mutations in *PKD1* (OMIM 173900), those with loss-of-function mutations have more severe disease and progress faster to ESRD than those with missense mutations[161,162]. Carriers of autosomal recessive disorders have generally been deemed suitable kidney donors, as individuals who are heterozygous for a recessive causal `allele` are not expected to develop the disease[160]. However, reports of milder, subclinical disease phenotypes among carriers suggest that these individuals are at higher risk of developing nephropathy than previously thought and therefore may warrant nephrological surveillance. For example, hepatorenal involvement (renal echogenicity and/or liver cysts) has been noted in obligate heterozygote parents of patients with ARPKD (OMIM 263200)[163], and mild renal acidification defects and nephrolithiasis have been observed in individuals who are heterozygous for mutations in *ATP6V1B1*, which are associated with distal renal tubular acidosis with deafness (OMIM 267300)[164]. Thus, additional study is needed to assess the long-term implications of carrier status for renal function, risk of developing kidney disease, and outcomes following kidney donation.

The available epidemiological and genetic studies support the use of genetic testing in nephropathies of unknown aetiology, particularly in the setting of a compelling family history of early-onset renal failure. The 2017 US Renal Data System report notes that in ~14% of adult and ~19% of paediatric patients with incident ESRD[18], the clinical diagnosis is “other” or “unknown”, and the European Renal Association–European Dialysis and Transplant Association[17] and Australia and New Zealand Dialysis and Transplant[16] registries report similar statistics. As renal biopsy is generally contraindicated in ESRD, genetic testing is a promising diagnostic tool, and case reports demonstrate the utility of genetic findings in this patient population[165–168]. For example, in a 12-year-old boy presumed to have nonsyndromic infantile-onset retinal dystrophy, identification of Senior–Loken syndrome 5 (OMIM 609254) led to early diagnosis of unrecognized CKD, enabling preplanned initiation of dialysis, appropriate donor selection for renal transplantation, and surveillance of at-risk family members[167]. Similarly, detection of *INF2*-mediated focal segmental glomerulosclerosis (OMIM 613237)[166] and *LMX1B* glomerulopathy (OMIM 161200)[165] in patients with familial ESRD of unknown origin helped to inform transplant prognosis and choice of donor, as these diagnoses are associated with low risk of disease recurrence and indicate genetic screening for candidate living related donors. Genetic diagnosis can also lead to targeted therapy and improved post-transplantation outcomes. For example, genetic testing of a 67-year-old woman with ESRD of unknown aetiology revealed adenine phosphoribosyltransferase (APRT) deficiency (OMIM 614723)[168]; the resulting initiation of xanthine dehydrogenase inhibitor therapy (allopurinol) prevented recurrence of crystalline nephropathy and allograft loss in this patient. Given the notable prevalence of kidney disease of unknown aetiology, even a modest diagnostic yield from genetic testing could have a large impact on clinical care.

The value of a genetic diagnosis

In addition to pinpointing the cause of disease, genetic diagnosis can inform clinical prognosis and guide patient management (TABLE 2). In general, patients and their physicians want a clear understanding of their disease, why it occurred, and how it will affect their health and medical care as well as that of their families[169]. As distinct pathophysiological processes can result in indistinguishable clinical presentations, the precise aetiology can remain unclear despite extensive history-taking and biochemical, imaging, and histopathological studies. This 'diagnostic odyssey' has substantial time, financial, and psychosocial costs for patients and their families and is a substantial burden on the health-care system. Thus, in addition to helping guide subsequent care, a genetic diagnosis holds substantial value through ending this process[27,69,170,171].

Importantly, translating genetic findings into improved patient care requires longitudinal studies of large cohorts of individuals with genetic diagnoses. Long-term follow-up of patients undergoing sequencing is needed to study the impact of genetic information on clinical management, health-care utilization, and outcomes. Rare disease referral networks, such as the European Reference Network for Rare Kidney Diseases[172,173], the National Institute of Health Rare Diseases Clinical Research Network[174], and the United Kingdom Kidney Research Consortium[175], will help to achieve this aim. Moreover, the knowledge generated will enable the development of best practice guidelines regarding diagnostic work-up and treatment of rare hereditary renal disorders. A genetic diagnosis can also enable referral for targeted clinical trials of therapies that might provide benefit in specific patient populations, such as microRNA inhibition for Alport syndrome[176] (OMIM 104200, 203780, 301050) and small-interfering-RNA blockade for primary hyperoxaluria type 1[177] (OMIM 259900). Disease-specific support groups can also help to direct patients to trials and other relevant clinical resources[178–180] and serve as key sources of psychosocial support for affected individuals and their families.

Finally, genetic-based stratification of clinical trials has the potential to prevent exposure to unnecessary risk of patients who are unlikely to benefit from an intervention while reducing confounders that may mask its benefit. For example, genetic testing could be used to exclude patients with hereditary forms of nephrotic syndrome, who do not tend to respond to steroid therapy[181], from a trial investigating the efficacy of a novel corticosteroid agent.

Clinical sequence interpretation

The aim of clinical sequence interpretation is to identify the genetic variant that is responsible for the phenotype of an individual patient. As the identified variant is used to guide subsequent care, it is critical that clinical sequence interpretation be highly accurate and reproducible[182]. However, the abundance of sequence variation in a typical human

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

genome and the vast search space provided by genome-wide testing results in a high risk of falsely ascribing causality to benign variants[183]. Although established guidelines for diagnostic interpretation exist[67,68,182], the process often remains time-consuming and highly subjective, requiring expert judgement at each step. A variety of online resources are available to aid geneticists and clinicians as they navigate the process of diagnostic sequence interpretation and application of genetic findings into clinical care (TABLE 3).

Diagnostic analysis is guided by the phenotype of the patient. Ordering clinicians are thus instructed to provide accurate and complete clinical information[182]; however, no standards currently exist for what information should be provided and/or who is qualified to give this information[184]. Moreover, even ostensibly specific clinical diagnoses may comprise a wide array of genetic and acquired disorders. For example, a diagnosis of primary FSGS could result from mutations in any one of >50 genes[124]; without further clinical information, such as age of onset, presence of extrarenal features, or pattern of inheritance, a geneticist would be hard-pressed to prioritize candidate variants during subsequent sequence interpretation. Referring clinicians should, therefore, provide detailed clinical information to the diagnostic laboratory and be prepared to discuss the patient in greater detail with the molecular geneticist in charge of sequence interpretation.

The first step in clinical sequence interpretation is to select the genes in which mutations can result in a phenotype compatible with the clinical presentation of the patient (FIG. 1). Attempts have been made to catalogue the genes that are associated with hereditary forms of nephropathy and classify them by their associated broad phenotype[1,4,185], but no systematic procedures or consensus guidelines exist regarding which genes should be evaluated for a given category of kidney disease. The choice is becoming increasingly complex owing to phenotypic expansions and is complicated because genes that are traditionally associated with nonrenal disorders can also present with nephropathy[186–188]. For example, mutations in *HNF4A* are classically implicated in maturity-onset diabetes of the young type 1 (MODY1; OMIM 125850) without renal involvement; however, the p.R76W `missense variant` has been noted in patients presenting with both Fanconi proximal tubular syndrome and MODY1 (REFS 186,187). Similarly, the identification of mutations in *FOXP1* in patients with CAKUT suggest that the phenotypic spectrum of these mutations encompasses CAKUT as well as intellectual disability (OMIM 613670)[188]. These situations require geneticists with domain expertise, who can recognize the causal connection between the mutation and the kidney phenotype. An additional challenge is the continuous need to assess the strength of gene–disease associations, as new genes continue to be rapidly identified[21] and additional research can cast doubt on those that have previously been implicated in a disease[189,190].

Interpretation at the variant level holds further complexity. Current guidelines evaluate a given variant in the context of the genetic architecture of the disease and the available literature. Lines of evidence include previous case reports, being in the same region and/or belonging to one of the functional types of variants previously noted as being causal for the disorder, having a population frequency compatible with that expected for the disease, and experimental and/or *in silico* support for a deleterious effect on protein function. Geneticists must examine the relevant observations for each of these criteria and combine them to arrive at an overall variant-level classification of pathogenic, likely pathogenic, likely benign, benign, or VUS[182].

Finally, the genetic findings must be assessed for concordance with the clinical presentation of the patient in addition to the genetic architecture of the disease. In some cases, this process is fairly straightforward; for example, as `haploinsufficiency` is recognized as the genetic mechanism of *HNF1B*-mediated disease[134] and *HNF1B* is highly intolerant of `loss-of-function variation`, a novel `nonsense variant` found in a patient with cystic renal disease and early-onset diabetes can be deemed likely causal. By contrast, a novel *HNF1B* missense variant, even if predicted to be deleterious based on multiple *in silico* algorithms, would require additional evidence to support pathogenicity. Additional work-up might involve parental testing to ascertain *de novo* status in sporadic cases, examining segregation if multiple affected family members are present, evaluating the patient for other features associated with the candidate diagnosis (for example, in the case of *HNF1B*-mediated disease, diabetes, gout, hypomagnesaemia, and genital abnormalities), and/or performing functional studies to model the effect of the variant on protein function. Such work-up is critical to clarify the clinical relevance of genetic findings, especially in the context of large multigene panels or genome-wide testing, where multiple candidate variants may be found[182,184,191]. Additional genetic and/or clinical testing can, however, involve substantial time and monetary costs[192] and, in some cases, might not be feasible (for example, blood relatives may be unavailable for genetic testing, preventing `variant phasing` or the determination of *de novo* status).

Population-wide `allele frequency` data have emerged as powerful first-line tools in clinical sequence interpretation[183]. The development of large public sequence databases has shed light on the spectrum of allele frequencies across populations of diverse ancestries[29,193] and has demonstrated that a large number of previously reported variants are unlikely to be pathogenic because they are present at frequencies exceeding the prevalence of the associated disease[193–196]. For example, in the Human Gene Mutation Database, the p.Ser487Leu variant in the *EYA1* gene is noted as causal for branchio-oto-renal syndrome (BOR; OMIM 113650) on the basis of two publications[197,198]. However, this variant has been reclassified by four independent diagnostic laboratories as a VUS or likely benign[199], and noted as a VUS in two publications[76,81], because of its high prevalence in the general population. The variant is present in 197 individuals in gnomAD (a large population control database), corresponding to a global frequency of 1 in 1,500, and is present in more than 1 in 1,000 Europeans. By contrast, BOR is estimated to have a prevalence of 1 in 40,000 and is thought to be fully penetrant[200]. With frequencies 28-fold and 50-fold higher than the total prevalence of BOR, the p.Ser487Leu variant is unlikely to be causal for such a rare, highly penetrant, autosomal dominant, monogenic disorder.

Thus, a great need exists for review of clinical variant databases using newly available population genetic data. In the meantime, variant interpretation will require time-intensive and subjective curation of the primary literature in the context of often limited knowledge of the prevalence, `penetrance`, and expressivity of a disease. In addition, many of the variants in clinical databases may have been classified by a single diagnostic laboratory; with no further explanation and/or supporting data, such findings have limited value.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Given the many layers of complexity, the existence of persistent interlaboratory and even inter-reviewer discordance in variant classification[201–204] despite adoption of fine-grained variant interpretation guidelines[182] is unsurprising. Frameworks for semiquantitative assessment of the clinical validity of gene–disease associations[205] and semiautomated clinical variant interpretation[206,207] have been proposed and might help to increase the reproducibility and efficiency of clinical sequence interpretation. However, these approaches still rely on subjective review of genetic and experimental evidence from the primary literature, and thus, the potential for divergent interpretation remains. Creation of consensus guidelines and quantitative standards will enable more objective and automated analysis at some steps, but as clinical sequence interpretation ultimately relies on clinical judgement, some degree of subjectivity will remain.

Applying genetic findings in the clinic

A genetic diagnosis provides a valuable answer but is only a starting point. For genetic findings to have clinical utility, they must be applied in the context of clinical care. This process is as complex as that of clinical sequence interpretation, and multiple barriers must be overcome to enable the promise of genomic medicine to be achieved. Key challenges in the implementation of genetic findings into clinical nephrology include return of results, physician education, sequence reanalysis, and the consideration of ELSIs[208–210].

Return of results

Clinical genetic testing is rapidly moving towards genome-wide assessment[32,35,63]. This expanded genetic scope increases diagnostic sensitivity but also has the potential to identify variants that are unrelated to the primary indication for testing. Such secondary findings must be considered with respect to their clinical validity and actionability[211]. Clinically valid findings include those that can be used to accurately predict that a patient will have the associated condition[212]; these encompass variants in genes for highly penetrant Mendelian diseases, pharmacogenomic variants that are informative regarding drug metabolism, and risk variants that affect susceptibility for a given condition. This category also includes clinically actionable variants, the detection of which would enable a physician to implement interventions that prevent or lessen the clinical consequences of the disease for which the variant confers increased risk. Clinically actionable variants have been recommended for return by both the ACMG[213,214] and the ESHG[215].

Currently, the ACMG advises returning known and expected pathogenic variants in 59 genes to patients regardless of their age or indication[214]. These genes encompass conditions deemed to be highly penetrant and actionable and predominantly consist of those that are associated with various hereditary forms of cancer and cardiovascular disease. Sequencing studies on large, unselected adult cohorts show that approximately 1–3% of the general population has a pathogenic mutation in one of these 59 genes[216,217]. Importantly, the ACMG actionable list includes genes that are associated with conditions relevant to renal medicine, such that the broadening use of genome sequencing may lead to additional nephrology consultations. These conditions include hereditary pheochromocytoma-paraganglioma syndrome (OMIM 168000, 601650, 605373, 115310), multiple endocrine

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

neoplasia (OMIM 131100, 171400, 162300), Wilms tumour (OMIM 194070), Fabry disease (OMIM 301500), Von Hippel–Landau syndrome (OMIM 193300), and tuberous sclerosis complex (OMIM 191100, 613254). Moreover, the detection of actionable mutations in other ACMG genes can impact nephrologic care, such as pretransplant defibrillator implantation in kidney transplant recipients with a *KCNQ1* mutation causal for long QT syndrome (OMIM 192500) or reduction of the dosage of immunosuppressive therapy in patients with mutations in genes that are associated with hereditary cancers, such *BRCA1*.

The ACMG encourages clinical specialists to nominate gene–disease pairs that they feel meet these actionability criteria as well as selected pharmacogenomic variants for medications that are commonly prescribed and/or associated with serious adverse events. Thus, there may be an opportunity to add nephrology-specific loci to the ACMG list, such as genes for highly penetrant and medically actionable hereditary nephropathies and pharmacogenomic variants that affect metabolism of medications commonly used in the care of patients with CKD[218,219]. Formation of multicentre interdisciplinary working groups and use of evidence-based frameworks to assess disease actionability[220] would greatly facilitate this effort.

Continuous review

As new genetic knowledge emerges, classifications shift. New genetic or experimental data may reclassify a mutation that was previously deemed diagnostic as a VUS or benign variant as pathogenic. Moreover, discovery of new genes may lead to a genetic diagnosis upon review of WES or WGS data from previously unsolved cases. As physicians use genetic diagnoses to guide care, such shifts are hugely important — altered classification of a variant can result in altered management of patients and their families. Yet no explicit standards for review of clinical genetic testing data currently exist, and the practice is rare, with a 2017 study reporting that only 1 of 21 laboratories surveyed routinely engaged in the practice[221].

In addition to questions regarding the optimal frequency and analytical methodology, continuous review of sequence data involves a multitude of ELSIs, including who should be responsible for requesting reanalysis, physician liability and duty to inform versus the right of patients not to know, and the psychosocial impact of recontact on patients and their families[222,223]. Given the scientific and ethical complexity and the many stake-holders involved, continuous review is not an easy issue to address or to create policies around; nevertheless, the issue is unavoidable as sequencing is increasingly incorporated into clinical practice. Comprehensive study of the clinical utility, cost-effectiveness, and psychosocial impact of continuous review and dialogue between relevant stakeholders, including patients, physicians, and genetics professionals, will help to ascertain its advantages and drawbacks and enable creation of formal guidelines regarding its implementation.

Genetic education and counselling

Under current guidelines for clinical genome-wide testing, the ordering clinician is expected to ensure informed consent by providing patients with comprehensive pretest counselling, including discussing the limitations of the test, the potential for secondary findings, and the complexity of the genetic interpretation[68,224]. The clinician is likewise expected to return

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

not only the primary results but also any secondary findings if the patient has opted to receive them. Investigations to date show that the majority of patients do opt to receive secondary findings[225], which necessitates a greater breadth and depth of genetics knowledge for the interpretation of their test results. However, there is currently a shortage of clinicians with adequate knowledge in genetics, genetic counsellors, and clinical geneticists who are capable of administering such comprehensive counselling[226,227]. Although genetics is being increasingly included in medical education, reports suggest that limited time and a lack of integration of genetics with clinical topics leaves students unprepared to apply genomics in patient care[228]. For example, a survey of recently graduated nephrology fellows noted that 65% felt that they had insufficient competence regarding genetic renal diseases, despite considering them important to their current clinical practice[229]. Among practising clinicians, no requirement currently exists for nongeneticists to have any specific knowledge or competencies in genetics[230], such that more senior clinicians may also feel unprepared.

Given the complexity of interpretation of genome-wide data and the many demands on their time, expecting nephrologists to act as genetic counsellors may be unreasonable. Rather, nephrologists should be expected to have a basic familiarity with genetics as well as more detailed knowledge of genetic forms of renal disease and to be aware of the general best practices regarding genetic testing. In the future, nephrogenetics may emerge as a superspecialty, similar to transplant or interventional nephrology. Genetic counsellors specialized in nephrology should become part of multidisciplinary teams fully integrated into the clinical setting, as has been successfully implemented in oncology[231] and cardiology[232]. In the absence of a geneticist in the clinical team, patients should be referred to genetic counsellors for counselling before and after testing. Efforts to provide genetics education to physicians, foster interaction between referring physicians and clinical testing laboratories, and create clinical decision support tools will help to achieve this aim and will facilitate implementation of clinical care based on genetic findings[233,234].

Ethical, legal, and social implications

The ELSI programme was founded as an integral part of the Human Genome Project with the understanding that genetic information can affect individuals, families, and society in a way that few other medical findings can. The historical role of human genetics in the development of eugenics and associated governmental policies such as compulsory sterilization invite caution as sequencing is introduced to general clinical practice. Here, we highlight some of the key ELSIs that must be considered when implementing genomic medicine in nephrology.

Participation in genetic research

In the landscape of genomic medicine, the boundaries between bench research and bedside care have become increasingly blurred. Participating in research can directly impact the clinical care and disease course of patients with hereditary nephropathies, for example, through enrolling in clinical trials such as the EARLY PRO-TECT trial, which is assessing the safety and efficacy of angiotensin-converting-enzyme inhibition in children with Alport syndrome[235,236]. In addition, researchers have begun to integrate patients' electronic health-

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

care records with their genetic data to study complex phenotypes such as resistant hypertension[237], and the promise of receiving medically actionable findings is a major incentive for many participants in genetic research[238,239].

The results generated through taking part in genetic studies differ markedly, however, from those that are returned from clinical diagnostic testing. A clinical genetic test aims to provide a genetic diagnosis for a patient within a defined time period, whereas sequencing performed in the context of genetic research aims to generate generalizable knowledge useful for future patients[240]. Although the latter approach might lead to diagnostic results, the time frame for their return is generally indefinite, and owing to the lack of regulation surrounding research-level genetic sequencing, putatively causal variants should be validated in a Clinical Laboratory Improvement Amendments (CLIA) certified laboratory before they can be returned in the context of clinical care[241,242]. Some have argued for the unification of research and clinical genetic testing, including requiring all research-level sequencing to be conducted in a CLIA laboratory environment to enable genetic discoveries to be rapidly implemented in patient care[243,244]. In the meantime, however, genetic counselling is needed to direct patients to the indicated sequencing modality given the clinical context (such as the urgency of a genetic diagnosis) and to ensure that they have appropriate expectations regarding what they may learn and in what time frame.

Minority populations

Substantial racial and ethnic disparities, which likely stem from both genetic and environmental factors, exist in kidney disease prevalence, severity, and treatment outcomes[245,246]. Importantly, knowledge of ancestry-specific alleles can have clinical utility for Mendelian nephropathies and for more common, complex forms of CKD. For example, founder mutations for autosomal recessive Alport syndrome (OMIM 203780)[247], Fanconi anaemia type C (OMIM 227645)[248], and Zellweger syndrome (OMIM 614866)[249] have been documented among Ashkenazi Jews, such that carrier screening may be beneficial in this population. With respect to more common, complex forms of CKD, the *APOL1* risk genotypes (BOX 1) substantially impact disease risk and clinical outcomes among patients of sub-Saharan African descent[250–252,287], and the sickle cell trait (*HBB* variant) may similarly confer increased CKD risk in this population[253,254].

Knowledge of population-specific allele frequencies can also help to prevent genetic misdiagnosis. Owing to a paucity of appropriate population controls, minority genomes have been frequently interpreted in the context of European sequence data, yielding an excess of reportedly pathogenic variants and VUS[255,256]. For example, in an analysis of genetic variants listed as pathogenic for hypertrophic cardiomyopathy, those with frequencies that were too high to be compatible with disease causality were significantly over-represented among healthy African Americans[256]. Importantly, these benign variants had been returned as diagnostic findings to African-American patients undergoing genetic testing for hypertrophic cardiomyopathy. Such false positive results can result in inaccurate understanding of risk status, medical mismanagement (for example, unnecessary implantation of a cardioverter–defibrillator), and substantial psychological distress for patients and their families. Thus, evaluation of sequence data in the context of a patient's

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

ancestry is critical to avert variant misinterpretation and the associated clinical perils. In addition, diverse population control databases are needed to address such disparities in genetic diagnostic testing among minority populations[257,258].

Genetic testing has been historically underutilized by minority populations, reflecting disparities in health-care access as well as negative perceptions regarding its application[259–261]. To ensure that the benefits of genomic nephrology are available to all patients with kidney disease, it is imperative that minority populations have equal access to participation in genetic research and to clinical genetic assessment. Ongoing efforts to identify and address existing barriers to participation, perform sequencing studies in diverse populations, and develop educational materials tailored to individuals of varied backgrounds will help to achieve these aims[262].

Resource-limited settings

In low-income and middle-income countries (LMICs), access to clinical screening and care, including genetic testing, can often be limited[263–265]. Thus, individuals with hereditary nephropathies are likely underdetected and cannot access appropriate care. Several strategies have been recommended to address these disparities[156,266]. Provision of logistical and technical support will help to build national registries, biobanks, and other infrastructure and to establish standards of care. Electronic modes of communication can enable international collaboration with global experts in rare nephropathies, helping local physicians to create and implement best practices for patients. Proposed approaches to implementing genetic testing in LMICs emphasize providing clinically useful and cost-effective services tailored to fit the needs of the given population. Strategies include targeted testing for founder variants in genes associated with medically actionable autosomal recessive nephropathies that have high prevalence in the region (for example, screening of the *MEFV* gene, associated with familial Mediterranean fever (OMIM 134610, 249100), in the southeast Mediterranean[267]) and focused NGS panels containing the genes that are most commonly mutated in the disorder (for example, a panel including *NPHS1, NPHS2*, and *WT1* for suspected hereditary early-onset nephrotic syndrome). Although WES or WGS are becoming the standard for work-up of undiagnosed disorders, the high cost and limited availability of these technologies are substantial barriers to their use in LMICs. Mendeliomes have been recommended as a means of providing broad assessment of the disease-associated genome at a feasible cost and have shown promising results as diagnostics in this context[268,269].

Organizations such as Human Heredity and Health in Africa (H3Africa) and the Mexico National Institute of Genomic Medicine (INMEGEN) have already begun to implement such strategies, which will help to make the benefits of genomic medicine available globally. The H3Africa Kidney Disease Research Network initiative centres upon a case–control study of 4,000 patients with CKD and 4,000 healthy controls across four sub-Saharan African nations[270]. The clinical diagnoses to be studied include SRNS, biopsy-proven glomerulopathy, CKD of unknown aetiology, and hereditary glomerulopathies. Patients will undergo comprehensive clinical work-up, including genetic testing, with long-term follow-up over a period of 5 years. Importantly, the H3Africa consortium is developing a new

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

genotyping array to ensure maximal coverage of the highly genetically diverse African genome[271]. In addition, the consortium aims to build the resources needed to support genomic nephrology in Africa through training physicians and researchers and establishing biorepositories and sequencing facilities. This initiative will provide critical data regarding the genetic architecture of kidney disease and help to build a strong foundation for genomic nephrology in the region, providing a model for subsequent global efforts.

Paediatric genetic testing

As early-onset CKD is enriched for genetic aetiologies, genetic testing has been advocated as a first-line diagnostic for paediatric nephropathy[4,156]. The discussion to date has overwhelmingly focused on the diagnostic value of genetic testing for this population; however, the ethical implications and psychosocial impact of genetic testing in minors must also be addressed. Important issues include the age at which consent for testing should be provided by the child rather than by the parents, how to balance the potential for early clinical intervention versus the right of the patient to an open future[272], how to explain the primary findings to the patient, and the potential impact of knowledge of a genetic condition on the psychoemotional development and health of the child.

With the advent of genome-wide testing, the question has also arisen of how secondary findings should be treated among paediatric patients. The majority of the genes designated by the ACMG for analysis of medically actionable secondary findings are associated with hereditary cancers and other adult-onset conditions. Thus, the value of returning such results to paediatric patients is debatable. Although this knowledge could resolve parental anxiety regarding risk status and inform family planning, it threatens the right of the patient to autonomy and confidentiality and could result in considerable psychological harm[273]. Moreover, no clear consensus guidelines currently exist regarding this topic. The American Academy of Pediatrics recommends deferring predictive testing for later-onset conditions until adulthood, whereas the ACMG advises that the decision be made on a per-family basis[274,275]. Matters become even more complex with adolescents, as they may want to decide independently of their parents or other family members and/or may not want to have their genetic findings shared with their families[276]. Longitudinal study of the medical and psychosocial impact of genetic testing on paediatric patients and their families will help to guide creation of best practice guidelines regarding these topics.

Legal protection

In 1997, the United Nations Educational, Scientific and Cultural Organization (UNESCO) Declaration on the Human Genome and Human Rights ruled that genetic information should not be used to infringe upon “human rights, fundamental freedoms, and human dignity”, establishing an international stance against `genetic discrimination` (GD). In Europe, the 1997 Convention on Human Rights and Biomedicine and the 2012 Charter of Fundamental Rights of the European Union protect individuals against GD; under the primacy of European law, these protections take precedence over any conflicting national legislation, and many European nations have also created their own anti-GD statutes[277,278]. Other nations, such as Australia, the US, and Canada, have also passed anti-GD legislation. By contrast, anti-GD protection remains scant in Latin America, Africa, and the Middle

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

East, with the exception of Mexico, Chile, Malawi, and Israel[278]. Anti-GD legislation is similarly sparse across Asia, despite the strong presence of genetic research and personalized medicine in many Asian nations; to date, only South Korea and Taiwan have created specific prohibitions[278,279].

Legal protection for individuals undergoing genetic testing has focused on regulating the use of genetic findings by insurers and employers, reflecting concerns that these parties will use genetic information to deny individuals employment and/or associated insurance benefits[277,280]. However, the degree of legal protection varies substantially between nations. The 2016 Council of Europe recommendation requires insurers to justify use of all "health-related personal data" and prohibits insurers from requiring individuals to undergo genetic testing for insurance purposes[281]. Similarly, in Israel, the Genetic Information Law of 2000 bars use of genetic information for employment and insurance purposes[282]. Other countries take a looser stance; in the UK, for instance, insurers may use genetic test results if approved by the government[283], and in Canada, insurers have pledged to not "use genetic test results for life insurance coverage of $250,000 or less" (REF. 284).

In the USA, the lack of a national health-care system has magnified concerns, as individuals usually obtain employment-based insurance. In an effort to protect individuals against GD, in 2008, the US government enacted the Genetic Information Nondiscrimination Act (GINA), which forbids employers and health insurers to request genetic information from individuals and/or discriminate against them on the basis of any available genetic information. These protections do not, however, extend to other types of insurance, including life and disability coverage, and have been threatened by subsequent legislative initiatives. For example, under the 2017 H.R.1313 bill, employers would be able to request that their employees undergo genetic testing and share their results as part of a workplace wellness programme. Employees who refuse could face increased yearly insurance premiums, a substantial financial penalty that some argue would effectively force individuals into participating, undermining their autonomy and privacy. Conversely, limited legal protection can also discourage individuals from choosing to participate in genomic research for fear that the data gathered would have to be shared and could be used to discriminate against them. Such fears could cause people to forgo potentially medically valuable tests and obstruct the clinical research needed to better understand and treat a variety of health conditions[285].

Conclusions

Genomic medicine aims to use genetic information about patients to inform their clinical care. CMA and NGS have revolutionized nephrology research, illuminating the molecular pathogenesis of a variety of genetic kidney diseases, and have great potential clinical utility across a wide range of indications. The remaining questions are how to fill in the substantial gaps in knowledge and how to translate what is currently known into personalized care. Strategies that may help to accomplish these aims include multicentre sequencing studies in large, diverse all-cause CKD cohorts; the establishment of expert working groups to create disease-specific standards for required pretest phenotypic information, genes assessed and variant interpretation; utilization of genetic stratification to better power clinical trials; and

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

the inclusion of geneticists and genetic counsellors in multidisciplinary care teams (TABLE 4).

While pursuing these efforts, it is imperative that we remain mindful of the limitations of our knowledge. Genetic testing does not give absolute answers, but rather provides a probabilistic biomarker, the meaning of which must be interpreted in the overall genomic and clinical context[286]. In many cases, the ‘one gene, one disease’ model does not apply owing to the presence of genetic and environmental modifiers. Thus, physicians and geneticists must incorporate diagnostic sequence interpretation with traditional tools such as clinical history and renal biopsy as well as with other sources of omic data, all of which can provide crucial insight into the genetic findings. Through considering each individual comprehensively in his or her own unique clinical context, genomic nephrology can deliver truly personalized care for patients with kidney disease.

Acknowledgments

The work of the authors was supported by grants from the US National Institutes of Health (1F30DK116473 (E.E.G.), 2R01DK080099, and 5U01HG008680 (A.G.G.)), and the American Society of Nephrology Foundation for Kidney Research Donald E. Wesson Research Fellowship (H.M.R.).

Glossary

Heritability

The proportion of interindividual variation in a given trait that is due to genetic factors.

Chromosomal microarray (CMA)

A technique to detect copy number variants by hybridizing a patient’s DNA to probes corresponding to various regions of the genome; the hybridization pattern for a given probe reflects the number of copies that the patient has of that genomic region.

Next-generation sequencing (NGS)

Simultaneous sequencing of multiple DNA segments; also known as massively parallel sequencing.

Genome

The entirety of an individual’s DNA. The genome is divided into smaller protein-coding segments called genes.

Genetic testing

The assessment of DNA sequence variation. Genetic testing can be performed at the level of a single variant, a gene, multiple genes, or the entire genome.

Genomic medicine

An emerging branch of medicine that uses information about an individual’s genome to inform their clinical care, including diagnosis, prognosis, and treatment.

Genetic diagnosis

The hereditary aetiology of a patient’s presentation, as identified by genetic testing.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Single-nucleotide variants (SNVs)

Changes of single bases (nucleotides) in a DNA sequence. SNVs can lead to an altered amino acid sequence in the encoded protein (nonsynonymous variants) or leave the sequence unchanged (synonymous variants).

Insertions or deletions

The gain or loss of bases in a DNA sequence, resulting in an altered amino acid sequence in the encoded protein.

Structural variants

Large (≥1 kb) DNA variants that include balanced (for example, inversions or reciprocal translocations) and imbalanced alterations (for example, copy number variants).

Sanger sequencing

A DNA sequencing method that uses labelled chain-terminating dideoxynucleotides to identify the nucleotides in the DNA strand being sequenced. This method generates a sequence chromatogram that can be analysed to detect genetic variants.

Targeted next-generation sequencing panels

Next-generation-sequencing-based analysis of a set of genes commonly associated with the patient’s clinically suspected phenotype.

Whole-exome sequencing (WES)

Next-generation-sequencing-based analysis of the exome — the protein-coding regions of the genome that contain the majority of known causal variants for Mendelian disorders.

Whole-genome sequencing (WGS)

Next-generation-sequencing-based analysis of the whole genome, including protein-coding and non-coding regions.

Karyotyping

A technique used to detect large genomic imbalances through visual inspection of stained chromosomes using a microscope at high magnification (× 1,000).

Copy number variants (CNVs)

Structural variants that results in gain or loss of DNA at the relevant locus.

Array comparative genomic hybridization

A type of chromosomal microarray in which patient and control DNA are labelled with different coloured fluorescent dyes and cohybridized to a single DNA probe in order to directly compare copy number at that genomic region.

Single-nucleotide polymorphism arrays

A type of chromosomal microarray in which a patient’s DNA is hybridized to DNA probes corresponding to single-nucleotide polymorphisms and the hybridization pattern is compared with previously analysed controls. This type of chromosomal microarray can detect a patient’s genotype in addition to copy number at a given genomic region.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Balanced chromosomal rearrangements

Chromosomal rearrangements that do not cause a net loss or gain of genetic material.

Sequencing coverage and depth

In this Review, sequencing coverage denotes the percentage of bases in the DNA region targeted by sequencing that is sequenced a given number of times. Sequencing depth refers to the average number of times that a given nucleotide is read in a set of DNA sequence reads. Higher coverage and depth means that more of the targeted genomic region has been sampled a greater number of times, increasing the accuracy of the resulting data.

Secondary findings

Genetic findings that are not related to the primary indication for testing; also called incidental findings.

Multiplex ligation-dependent probe amplification

A technique in which patient DNA is hybridized to two oligonucleotide probes, corresponding to the 5′ and 3′ ends of the DNA, which are then ligated and PCR-amplified using a fluorescently labelled primer. The resulting PCR products are size-separated using capillary electrophoresis, and the fluorescent signal intensity is compared between the probe and the patient’s DNA to determine copy number at that region. In addition to identifying copy number variants, this technique can detect mosaicism for a copy number variant and DNA methylation status.

Variants of uncertain significance (VUSs)

Genetic variants that have an unclear association with a given disorder owing to insufficient or conflicting evidence.

Phenotypic expansions

Phenotypic expansions occur when mutations in a gene that is classically associated with one phenotype are demonstrated to cause another clinically distinct phenotype.

Allele

Within each chromosome, the DNA sequence at a given region can vary; these variants are alleles.

Missense variant

Single-nucleotide variant that leads to the replacement of the amino acid normally encoded with another amino acid.

Haploinsufficiency

The state that arises when one copy of a gene is deleted or otherwise inactivated and the single remaining copy is insufficient to produce the amount of gene product needed to maintain normal function, leading to an abnormal (disease) phenotype.

Loss-of-function variation

DNA sequence alteration that leads to a protein with severely reduced or no function. Genetic variants that result in a prematurely truncated protein, such as nonsense variants, generally cause loss of function; however, missense variants can also have this effect.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Nonsense variant

Single-nucleotide variant that leads to the replacement of the amino acid normally encoded with a stop codon, leading to a prematurely truncated protein.

Variant phasing

Determining whether two variants in an individual's genome are both on the same copy of the gene (in cis) or on different copies of the genes (in trans) by use of parental testing. If two variants in a gene associated with a recessive disorder are in *trans*, they are more likely to be causal for the disorder, as they will impact both copies of the gene.

Allele frequency

The incidence of an allele in a population. Allele frequency is calculated by dividing the number of times that the allele is found by the total number of chromosomes. The allele frequency can be used to assess the rarity of a certain allele to help ascertain its pathogenicity during clinical sequence interpretation.

Penetrance

The proportion of individuals with a certain genetic variant who display the phenotype that is associated with this variant.

Genetic discrimination (GD)

Differential treatment of individuals on the basis of their genetic information.

References

1. Devuyst O, et al. Rare inherited kidney diseases: challenges, opportunities, and perspectives. Lancet. 2014; 383:1844–1859. This review provides a comprehensive overview of the major Mendelian forms of CKD and examines the existing challenges to and future opportunities for effective clinical detection and management. [PubMed: 24856029]
2. Jha V, et al. Chronic kidney disease: global dimension and perspectives. Lancet. 2013; 382:260–272. [PubMed: 23727169]
3. Wuhl E, et al. Renal replacement therapy for rare diseases affecting the kidney: an analysis of the ERA-EDTA Registry. Nephrol. Dial. Transplant. 2014; 29(Suppl. 4):iv1–iv8. [PubMed: 25165174]
4. Vivante A, Hildebrandt F. Exploring the genetic basis of early-onset chronic kidney disease. Nat. Rev. Nephrol. 2016; 12:133–146. [PubMed: 26750453]
5. Ingelfinger JR, Kalantar-Zadeh K, Schaefer F. World Kidney Day Steering Committee. World Kidney Day 2016: averting the legacy of kidney disease — focus on childhood. Pediatr. Nephrol. 2016; 31:343–348. [PubMed: 26884120]
6. Arpegard J, et al. Comparison of heritability of Cystatin C^- and creatinine-based estimates of kidney function and their relation to heritability of cardiovascular disease. J. Am. Heart Assoc. 2015; 4:e001467. [PubMed: 25589536]
7. Fox CS, et al. Genomewide linkage analysis to serum creatinine, GFR, and creatinine clearance in a community-based population: the Framingham Heart Study. J. Am. Soc. Nephrol. 2004; 15:2457–2461. [PubMed: 15339995]
8. Moulin F, et al. A population-based approach to assess the heritability and distribution of renal handling of electrolytes. Kidney Int. 2017; 92:1536–1543. [PubMed: 28888328]
9. Wilmot B, et al. Heritability of serum sodium concentration: evidence for sex- and ethnic-specific effects. Physiol. Genomics. 2012; 44:220–228. [PubMed: 22186255]
10. Lieske JC, Turner ST, Edeh SN, Smith JA, Kardia SL. Heritability of urinary traits that contribute to nephrolithiasis. Clin. J. Am. Soc. Nephrol. 2014; 9:943–950. [PubMed: 24578335]

11. Skrunes R, Svarstad E, Reisaeter AV, Vikse BE. Familial clustering of ESRD in the Norwegian population. Clin. J. Am. Soc. Nephrol. 2014; 9:1692–1700. [PubMed: 25092600]
12. Connaughton DM, et al. The Irish Kidney Gene Project — prevalence of family history in patients with kidney disease in Ireland. Nephron. 2015; 130:293–301. [PubMed: 26202451]
13. McClellan WM, et al. Individuals with a family history of ESRD are a high-risk population for CKD: implications for targeted surveillance and intervention activities. Am. J. Kidney Dis. 2009; 53:S100–S106. [PubMed: 19231753]
14. Joly D, Beroud C, Grunfeld JP. Rare inherited disorders with renal involvement-approach to the patient. Kidney Int. 2015; 87:901–908. [PubMed: 25651365]
15. Liapis H, Gaut JP. The renal biopsy in the genomic era. Pediatr. Nephrol. 2013; 28:1207–1219. [PubMed: 23179194]
16. Australia and New Zealand Dialysis and Transplant Registry. Annual ANZDATA Report. ANZDATA Registry, South Adelaide; Australia: 2016.
17. European Renal Association–European Dialysis and Transplant Association. ERA-EDTA Annual Report 2015. ERA-EDTA Registry. 2017. https://www.era-edta-reg.org/files/annualreports/pdf/AnnRep2015.pdf
18. United States Renal Data System. USRDS annual data report: epidemiology of kidney disease in the United States. National Institutes of Health; 2017.
19. James MT, Hemmelgarn BR, Tonelli M. Early recognition and prevention of chronic kidney disease. Lancet. 2010; 375:1296–1309. [PubMed: 20382326]
20. KDIGO. KDIGO 2012 clinical practice guideline for the evaluation and management of chronic kidney disease. Kidney Int. Suppl. 2013; 3:19–62.
21. Chong JX, et al. The genetic basis of mendelian phenotypes: discoveries, challenges, and opportunities. Am. J. Hum. Genet. 2015; 97:199–215. [PubMed: 26166479]
22. Smith LD, Willig LK, Kingsmore SF. Whole-exome sequencing and whole-genome sequencing in critically ill neonates suspected to have single-gene disorders. Cold Spring Harb. Perspect. Med. 2015; 6:a023168. [PubMed: 26684335]
23. Miller DT, et al. Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. Am. J. Hum. Genet. 2010; 86:749–764. [PubMed: 20466091]
24. Guttmacher AE, Collins FS. Genomic medicine — a primer. N. Engl. J. Med. 2002; 347:1512–1520. [PubMed: 12421895]
25. Dixon-Salazar TJ, et al. Exome sequencing can improve diagnosis and alter patient management. Sci. Transl Med. 2012; 4:138ra178.
26. Lee H, et al. Clinical exome sequencing for genetic identification of rare Mendelian disorders. JAMA. 2014; 312:1880–1887. [PubMed: 25326637]
27. Valencia CA, et al. Clinical impact and cost-effectiveness of whole exome sequencing as a diagnostic tool: a Pediatric Center's Experience. Front. Pediatr. 2015; 3:67. [PubMed: 26284228]
28. Stokman MF, et al. The expanding phenotypic spectra of kidney diseases: insights from genetic studies. Nat. Rev. Nephrol. 2016; 12:472–483. [PubMed: 27374918]
29. Genomes Project Consortium *et al*. A global reference for human genetic variation. Nature. 2015; 526:68–74. [PubMed: 26432245]
30. Online Mendelian Inheritance in Man. OMIM Gene Map Statistics. OMIM. 2017. https://www.omim.org/statistics/geneMap
31. Goldstein DB, et al. Sequencing studies in human genetics: design and interpretation. Nat. Rev. Genet. 2013; 14:460–470. [PubMed: 23752795]
32. Katsanis SH, Katsanis N. Molecular genetic testing and the future of clinical genomics. Nat. Rev. Genet. 2013; 14:415–426. [PubMed: 23681062]
33. Rehm HL. Disease-targeted sequencing: a cornerstone in the clinic. Nat. Rev. Genet. 2013; 14:295–300. [PubMed: 23478348]
34. Petersen BS, Fredrich B, Hoeppner MP, Ellinghaus D, Franke A. Opportunities and challenges of whole-genome and -exome sequencing. BMC Genet. 2017; 18:14. [PubMed: 28193154]

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

35. Xue Y, Ankala A, Wilcox WR, Hegde MR. Solving the molecular diagnostic testing conundrum for Mendelian disorders in the era of next-generation sequencing: single-gene, gene panel, or exome/genome sequencing. Genet. Med. 2015; 17:444–451. This review compares the major diagnostic genetic testing modalities and provides a generalized algorithm to help select the most appropriate test given a patient’s clinical presentation. [PubMed: 25232854]
36. Carvalho CM, Lupski JR. Mechanisms underlying structural variant formation in genomic disorders. Nat. Rev. Genet. 2016; 17:224–238. [PubMed: 26924765]
37. Watson CT, Marques-Bonet T, Sharp AJ, Mefford HC. The genetics of microdeletion and microduplication syndromes: an update. Annu. Rev. Genomics Hum. Genet. 2014; 15:215–244. [PubMed: 24773319]
38. Carter NP. Methods and strategies for analyzing copy number variation using DNA microarrays. Nat. Genet. 2007; 39:S16–S21. [PubMed: 17597776]
39. Schaaf CP, Wiszniewska J, Beaudet AL. Copy number and SNP arrays in clinical diagnostics. Annu. Rev. Genomics Hum. Genet. 2011; 12:25–51. [PubMed: 21801020]
40. Vermeesch JR, Brady PD, Sanlaville D, Kok K, Hastings RJ. Genome-wide arrays: quality criteria and platforms to be used in routine diagnostics. Hum. Mutat. 2012; 33:906–915. [PubMed: 22415865]
41. Kearney HM, et al. American College of Medical Genetics standards and guidelines for interpretation and reporting of postnatal constitutional copy number variants. Genet. Med. 2011; 13:680–685. [PubMed: 21681106]
42. Reddy UM, et al. Karyotype versus microarray testing for genetic abnormalities after stillbirth. N. Engl. J. Med. 2012; 367:2185–2193. [PubMed: 23215556]
43. Wapner RJ, et al. Chromosomal microarray versus karyotyping for prenatal diagnosis. N. Engl. J. Med. 2012; 367:2175–2184. [PubMed: 23215555]
44. South ST, et al. ACMG Standards and Guidelines for constitutional cytogenomic microarray analysis, including postnatal and prenatal applications: revision 2013. Genet. Med. 2013; 15:901–909. [PubMed: 24071793]
45. Pinto D, et al. Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. Nat. Biotechnol. 2011; 29:512–520. [PubMed: 21552272]
46. Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. Nat. Rev. Genet. 2011; 12:363–376. [PubMed: 21358748]
47. Harambat J, van Stralen KJ, Kim JJ, Tizard EJ. Epidemiology of chronic kidney disease in children. Pediatr. Nephrol. 2012; 27:363–373. [PubMed: 21713524]
48. Nicolaou N, Renkema KY, Bongers EM, Giles RH, Knoers NV. Genetic, environmental, and epigenetic factors involved in CAKUT. Nat. Rev. Nephrol. 2015; 11:720–731. [PubMed: 26281895]
49. Weber S, et al. Mapping candidate regions and genes for congenital anomalies of the kidneys and urinary tract (CAKUT) by array-based comparative genomic hybridization. Nephrol. Dial. Transplant. 2011; 26:136–143. [PubMed: 20605837]
50. Sanna-Cherchi S, et al. Copy-number disorders are a common cause of congenital kidney malformations. Am. J. Hum. Genet. 2012; 91:987–997. [PubMed: 23159250]
51. Tammimies K, et al. Molecular diagnostic yield of chromosomal microarray analysis and whole-exome sequencing in children with autism spectrum disorder. JAMA. 2015; 314:895–903. [PubMed: 26325558]
52. Caruana G, et al. Copy-number variation associated with congenital anomalies of the kidney and urinary tract. Pediatr. Nephrol. 2015; 30:487–495. [PubMed: 25270717]
53. Faure A, et al. DNA copy number variants: a potentially useful predictor of early onset renal failure in boys with posterior urethral valves. J. Pediatr. Urol. 2016; 12:227.e1–227.e7. [PubMed: 27160979]
54. Westland R, et al. Copy number variation analysis identifies novel CAKUT candidate genes in children with a solitary functioning kidney. Kidney Int. 2015; 88:1402–1410. [PubMed: 26352300]
55. Verbitsky M, et al. Genomic imbalances in pediatric patients with chronic kidney disease. J. Clin. Invest. 2015; 125:2171–2178. [PubMed: 25893603]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

56. Fu F, et al. Prenatal diagnosis of fetal multicystic dysplastic kidney via high-resolution whole-genome array. Nephrol. Dial. Transplant. 2016; 31:1693–1698. [PubMed: 26932690]
57. Moreira JM, Bouissou Morais Soares CM, Teixeira AL, Simoes ESAC, Kummer AM. Anxiety, depression, resilience and quality of life in children and adolescents with pre-dialysis chronic kidney disease. Pediatr. Nephrol. 2015; 30:2153–2162. [PubMed: 26210984]
58. Hooper SR, et al. Neurocognitive functioning of children and adolescents with mild-to-moderate chronic kidney disease. Clin. J. Am. Soc. Nephrol. 2011; 6:1824–1830. [PubMed: 21737850]
59. Ruebner RL, et al. Neurocognitive dysfunction in children, adolescents, and young adults with CKD. Am. J. Kidney Dis. 2016; 67:567–575. [PubMed: 26476795]
60. Verbitsky, M., et al. Genomic disorders and neurocognitive impairment in pediatric CKD. J. Am. Soc. Nephrol. 2017. http://dx.doi.org/10.1681/ASN.2016101108 This article shows that paediatric patients with CKD owing to genomic disorders have poorer neurocognitive function than those without independent of the severity of renal dysfunction. These results suggest that for some patients, neurocognitive impairment and nephropathy have a common genetic basis, informing clinical expectations and management.
61. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat. Rev. Genet. 2016; 17:333–351. This Review provides an excellent technical summary of existing and emerging NGS-based approaches for genomic sequencing. [PubMed: 27184599]
62. Levy SE, Myers RM. Advancements in next-generation sequencing. Annu. Rev. Genomics Hum. Genet. 2016; 17:95–115. [PubMed: 27362342]
63. Rehm HL. Evolving health care through personal genomics. Nat. Rev. Genet. 2017; 18:259–267. This Review highlights the different uses of genetic testing to achieve personalized health care in a range of clinical contexts and discusses key challenges to broadly implementing genomics into everyday medical practice. [PubMed: 28138143]
64. Ashley EA. Towards precision medicine. Nat. Rev. Genet. 2016; 17:507–522. [PubMed: 27528417]
65. Chakravorty S, Hegde M. Gene and variant annotation for mendelian disorders in the era of advanced sequencing technologies. Annu. Rev. Genomics Hum. Genet. 2017; 18:229–256. [PubMed: 28415856]
66. Rehm HL, et al. ACMG clinical laboratory standards for next-generation sequencing. Genet. Med. 2013; 15:733–747. [PubMed: 23887774]
67. Wallis, Y., et al. Practice Guidelines for the Evaluation of Pathogenicity and the Reporting of Sequence Variants in Clinical Molecular Genetics. Association for Clinical Genetic Science; 2013.
68. Matthijs G, et al. Guidelines for diagnostic next-generation sequencing. Eur. J. Hum. Genet. 2016; 24:2–5. [PubMed: 26508566]
69. Shashi V, et al. The utility of the traditional medical genetics diagnostic evaluation in the context of next-generation sequencing for undiagnosed genetic disorders. Genet. Med. 2014; 16:176–182. [PubMed: 23928913]
70. Sadowski CE, et al. A single-gene cause in 29.5% of cases of steroid-resistant nephrotic syndrome. J. Am. Soc. Nephrol. 2015; 26:1279–1289. [PubMed: 25349199]
71. McCarthy HJ, et al. Simultaneous sequencing of 24 genes associated with steroid-resistant nephrotic syndrome. Clin. J. Am. Soc. Nephrol. 2013; 8:637–648. [PubMed: 23349334]
72. Braun DA, et al. Prevalence of monogenic causes in pediatric patients with nephrolithiasis or nephrocalcinosis. Clin. J. Am. Soc. Nephrol. 2016; 11:664–672. [PubMed: 26787776]
73. Halbritter J, et al. Fourteen monogenic genes account for 15% of nephrolithiasis/nephrocalcinosis. J. Am. Soc. Nephrol. 2015; 26:543–551. [PubMed: 25296721]
74. Schueler M, et al. Large-scale targeted sequencing comparison highlights extreme genetic heterogeneity in nephronophthisis-related ciliopathies. J. Med. Genet. 2016; 53:208–214. [PubMed: 26673778]
75. Halbritter J, et al. High-throughput mutation analysis in patients with a nephronophthisis-associated ciliopathy applying multiplexed barcoded array-based PCR amplification and next-generation sequencing. J. Med. Genet. 2012; 49:756–767. [PubMed: 23188109]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

76. Hwang DY, et al. Mutations in 12 known dominant disease-causing genes clarify many congenital anomalies of the kidney and urinary tract. Kidney Int. 2014; 85:1429–1433. [PubMed: 24429398]
77. Kohl S, et al. Mild recessive mutations in six Fraser syndrome-related genes cause isolated congenital anomalies of the kidney and urinary tract. J. Am. Soc. Nephrol. 2014; 25:1917–1922. [PubMed: 24700879]
78. Gross O, et al. Advances and unmet needs in genetic, basic and clinical science in Alport syndrome: report from the 2015 International Workshop on Alport Syndrome. Nephrol. Dial. Transplant. 2017; 32:916–924. [PubMed: 27190345]
79. Savige J, et al. Expert guidelines for the management of Alport syndrome and thin basement membrane nephropathy. J. Am. Soc. Nephrol. 2013; 24:364–375. [PubMed: 23349312]
80. Moriniere V, et al. Improving mutation screening in familial hematuric nephropathies through next generation sequencing. J. Am. Soc. Nephrol. 2014; 25:2740–2751. [PubMed: 24854265]
81. Nicolaou N, et al. Prioritization and burden analysis of rare variants in 208 candidate genes suggest they do not play a major role in CAKUT. Kidney Int. 2016; 89:476–486. [PubMed: 26489027]
82. Heidet L, et al. Targeted exome sequencing identifies *PBX1* as involved in monogenic congenital anomalies of the kidney and urinary tract. J. Am. Soc. Nephrol. 2017; 28:2901–2914. [PubMed: 28566479]
83. Delio M, et al. Development of a targeted multi-disorder high-throughput sequencing assay for the effective identification of disease-causing variants. PLoS ONE. 2015; 10:e0133742. [PubMed: 26214305]
84. Saudi Mendeliome G. Comprehensive gene panels provide advantages over clinical exome sequencing for Mendelian diseases. Genome Biol. 2015; 16:134. [PubMed: 26112015]
85. Zemojtel T, et al. Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome. Sci. Transl Med. 2014; 6:252ra123.
86. Bamshad MJ, et al. Exome sequencing as a tool for Mendelian disease gene discovery. Nat. Rev. Genet. 2011; 12:745–755. [PubMed: 21946919]
87. Gilissen C, Hoischen A, Brunner HG, Veltman JA. Disease gene identification strategies for exome sequencing. Eur. J. Hum. Genet. 2012; 20:490–497. [PubMed: 22258526]
88. Eldomery MK, et al. Lessons learned from additional research analyses of unsolved clinical exome cases. Genome Med. 2017; 9:26. [PubMed: 28327206]
89. Wenger AM, Guturu H, Bernstein JA, Bejerano G. Systematic reanalysis of clinical exome data yields additional diagnoses: implications for providers. Genet. Med. 2016; 19:209–214. This study identified newly diagnostic mutations in 10% of patients with initially nondiagnostic WES results, highlighting the importance of regular sequence reanalysis in clinical testing. [PubMed: 27441994]
90. Need AC, Shashi V, Schoch K, Petrovski S, Goldstein DB. The importance of dynamic re-analysis in diagnostic whole exome sequencing. J. Med. Genet. 2017; 54:155–156. [PubMed: 27899421]
91. Bowling KM, et al. Genomic diagnosis for children with intellectual disability and/or developmental delay. Genome Med. 2017; 9:43. [PubMed: 28554332]
92. LaDuca H, et al. Exome sequencing covers >98% of mutations identified on targeted next generation sequencing panels. PLoS ONE. 2017; 12:e0170843. [PubMed: 28152038]
93. Mandelker D, et al. Navigating highly homologous genes in a molecular diagnostic setting: a resource for clinical next-generation sequencing. Genet. Med. 2016; 18:1282–1289. [PubMed: 27228465]
94. Park JY, et al. Clinical exome performance for reporting secondary genetic findings. Clin. Chem. 2015; 61:213–220. [PubMed: 25414276]
95. Botstein D, Risch N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nat. Genet. 2003; 33(Suppl):228–237. [PubMed: 12610532]
96. Cooper DN, et al. Genes, mutations, and human inherited disease at the dawn of the age of personalized genomics. Hum. Mutat. 2010; 31:631–655. [PubMed: 20506564]
97. Boycott KM, Vanstone MR, Bulman DE, MacKenzie AE. Rare-disease genetics in the era of next-generation sequencing: discovery to translation. Nat. Rev. Genet. 2013; 14:681–691. [PubMed: 23999272]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

98. Ku CS, Naidoo N, Pawitan Y. Revisiting Mendelian disorders through exome sequencing. Hum. Genet. 2011; 129:351–370. [PubMed: 21331778]

99. Taylor JC, et al. Factors influencing success of clinical genome sequencing across a broad spectrum of disorders. Nat. Genet. 2015; 47:717–726. [PubMed: 25985138]

100. Lupski JR, et al. Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. N. Engl. J. Med. 2010; 362:1181–1191. [PubMed: 20220177]

101. Saunders CJ, et al. Rapid whole-genome sequencing for genetic disease diagnosis in neonatal intensive care units. Sci. Transl Med. 2012; 4:154ra135.

102. Yuen RK, et al. Whole-genome sequencing of quartet families with autism spectrum disorder. Nat. Med. 2015; 21:185–191. [PubMed: 25621899]

103. Mele C, et al. Characterization of a new DGKE intronic mutation in genetically unsolved cases of familial atypical hemolytic uremic syndrome. Clin. J. Am. Soc. Nephrol. 2015; 10:1011–1019. [PubMed: 25854283]

104. King K, Flinter FA, Nihalani V, Green PM. Unusual deep intronic mutations in the COL4A5 gene cause X linked Alport syndrome. Hum. Genet. 2002; 111:548–554. [PubMed: 12436246]

105. Carroll C, Hunley TE, Guo Y, Cortez D. A novel splice site mutation in SMARCAL1 results in aberrant exon definition in a child with Schimke immunoosseous dysplasia. Am. J. Med. Genet. A. 2015; 167A:2260–2264. [PubMed: 25943327]

106. Lo YF, et al. Recurrent deep intronic mutations in the SLC12A3 gene responsible for Gitelman’s syndrome. Clin. J. Am. Soc. Nephrol. 2011; 6:630–639. [PubMed: 21051746]

107. Cabezas OR, et al. Polycystic kidney disease with hyperinsulinemic hypoglycemia caused by a promoter mutation in phosphomannomutase 2. J. Am. Soc. Nephrol. 2017; 28:2529–2539. [PubMed: 28373276]

108. Belkadi A, et al. Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. Proc. Natl Acad. Sci. USA. 2015; 112:5473–5478. [PubMed: 25827230]

109. Lelieveld SH, Spielmann M, Mundlos S, Veltman JA, Gilissen C. Comparison of exome and genome sequencing technologies for the complete capture of protein-coding regions. Hum. Mutat. 2015; 36:815–822. [PubMed: 25973577]

110. Mallawaarachchi AC, et al. Whole-genome sequencing overcomes pseudogene homology to diagnose autosomal dominant polycystic kidney disease. Eur. J. Hum. Genet. 2016; 24:1584–1590. [PubMed: 27165007]

111. Watson CM, et al. Enhanced diagnostic yield in Meckel–Gruber and Joubert syndrome through exome sequencing supplemented with split-read mapping. BMC Med. Genet. 2016; 17:1. [PubMed: 26729329]

112. Carss KJ, et al. Comprehensive rare variant analysis via whole-genome sequencing to determine the molecular pathology of inherited retinal disease. Am. J. Hum. Genet. 2017; 100:75–90. [PubMed: 28041643]

113. Gilissen C, et al. Genome sequencing identifies major causes of severe intellectual disability. Nature. 2014; 511:344–347. [PubMed: 24896178]

114. Stavropoulos DJ, et al. Whole genome sequencing expands diagnostic utility and improves clinical management in pediatric medicine. NPJ Genomics Med. 2016; 1:15012.

115. Kirby A, et al. Mutations causing medullary cystic kidney disease type 1 lie in a large VNTR in MUC1 missed by massively parallel sequencing. Nat. Genet. 2013; 45:299–303. [PubMed: 23396133]

116. Blumenstiel B, et al. Development and validation of a mass spectrometry-based assay for the molecular diagnosis of mucin-1 kidney disease. J. Mol. Diagn. 2016; 18:566–571. [PubMed: 27157321]

117. Dewey FE, et al. Clinical interpretation and implications of whole-genome sequencing. JAMA. 2014; 311:1035–1045. This article demonstrates the technical and interpretative challenges of using WGS in a clinical setting, highlighting key priorities to be addressed in order for this modality to be employed in patient care. [PubMed: 24618965]

118. Chrystoja CC, Diamandis EP. Whole genome sequencing as a diagnostic test: challenges and opportunities. Clin. Chem. 2014; 60:724–733. [PubMed: 24227285]

119. Braun DA, Hildebrandt F. Ciliopathies. Cold Spring Harb. Perspect. Biol. 2017; 9:a028191. [PubMed: 27793968]

120. Renkema KY, Stokman MF, Giles RH, Knoers NV. Next-generation sequencing for research and diagnostics in kidney disease. Nat. Rev. Nephrol. 2014; 10:433–444. [PubMed: 24914583]

121. Halbritter J, et al. Identification of 99 novel mutations in a worldwide cohort of 1,056 patients with a nephronophthisis-related ciliopathy. Hum. Genet. 2013; 132:865–884. [PubMed: 23559409]

122. Braun DA, et al. Whole exome sequencing identifies causative mutations in the majority of consanguineous or familial cases with childhood-onset increased renal echogenicity. Kidney Int. 2016; 89:468–475. [PubMed: 26489029]

123. Gee HY, et al. Whole-exome resequencing distinguishes cystic kidney diseases from phenocopies in renal ciliopathies. Kidney Int. 2014; 85:880–887. [PubMed: 24257694]

124. Bierzynska A, et al. Genomic and clinical profiling of a national nephrotic syndrome cohort advocates a precision medicine approach to disease management. Kidney Int. 2017; 91:937–947. [PubMed: 28117080]

125. Heyer CM, et al. Predicted mutation strength of nontruncating PKD1 mutations aids genotype–phenotype correlations in autosomal dominant polycystic kidney disease. J. Am. Soc. Nephrol. 2016; 27:2872–2884. [PubMed: 26823553]

126. Audrezet MP, et al. Autosomal dominant polycystic kidney disease: comprehensive mutation analysis of PKD1 and PKD2 in 700 unrelated patients. Hum. Mutat. 2012; 33:1239–1250. [PubMed: 22508176]

127. Gunay-Aygun M, et al. PKHD1 sequence variations in 78 children and adults with autosomal recessive polycystic kidney disease and congenital hepatic fibrosis. Mol. Genet. Metab. 2010; 99:160–173. [PubMed: 19914852]

128. Krall P, et al. Cost-effective PKHD1 genetic testing for autosomal recessive polycystic kidney disease. Pediatr. Nephrol. 2014; 29:223–234. [PubMed: 24162162]

129. Porath B, et al. Mutations in GANAB, encoding the glucosidase IIalpha subunit, cause autosomal-dominant polycystic kidney and liver disease. Am. J. Hum. Genet. 2016; 98:1193–1207. [PubMed: 27259053]

130. Lu H, et al. Mutations in DZIP1L, which encodes a ciliary-transition-zone protein, cause autosomal recessive polycystic kidney disease. Nat. Genet. 2017; 49:1025–1034. [PubMed: 28530676]

131. Gast C, et al. Collagen (COL4A) mutations are the most frequent mutations underlying adult focal segmental glomerulosclerosis. Nephrol. Dial. Transplant. 2016; 31:961–970. [PubMed: 26346198]

132. Malone AF, et al. Rare hereditary COL4A3/COL4A4 variants may be mistaken for familial focal segmental glomerulosclerosis. Kidney Int. 2014; 86:1253–1259. [PubMed: 25229338]

133. Verhave JC, Bech AP, Wetzels JF, Nijenhuis T. Hepatocyte nuclear factor 1beta-associated kidney disease: more than renal cysts and diabetes. J. Am. Soc. Nephrol. 2016; 27:345–353. [PubMed: 26319241]

134. Clissold RL, Hamilton AJ, Hattersley AT, Ellard S, Bingham C. HNF1B-associated renal and extra-renal disease-an expanding clinical spectrum. Nat. Rev. Nephrol. 2015; 11:102–112. [PubMed: 25536396]

135. Barua M, et al. Mutations in PAX2 associate with adult-onset FSGS. J. Am. Soc. Nephrol. 2014; 25:1942–1953. [PubMed: 24676634]

136. Huynh Cong E, et al. A homozygous missense mutation in the ciliary gene TTC21B causes familial FSGS. J. Am. Soc. Nephrol. 2014; 25:2435–2443. [PubMed: 24876116]

137. Bullich G, et al. Contribution of the TTC21B gene to glomerular and cystic kidney diseases. Nephrol. Dial. Transplant. 2017; 32:151–156. [PubMed: 26940125]

138. Scolari F, Izzi C, Ghiggeri GM. Uromodulin: from monogenic to multifactorial diseases. Nephrol. Dial. Transplant. 2015; 30:1250–1256. [PubMed: 25228753]

139. Besbas N, Ozaltin F, Jeck N, Seyberth H, Ludwig M. CLCN5 mutation (R347X) associated with hypokalaemic metabolic alkalosis in a Turkish child: an unusual presentation of Dent’s disease. Nephrol. Dial. Transplant. 2005; 20:1476–1479. [PubMed: 15814539]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

140. Okamoto T, Tajima T, Hirayama T, Sasaki S. A patient with Dent disease and features of Bartter syndrome caused by a novel mutation of CLCN5. Eur. J. Pediatr. 2012; 171:401–404. [PubMed: 21932010]

141. Fervenza FC. A patient with nephrotic-range proteinuria and focal global glomerulosclerosis. Clin. J. Am. Soc. Nephrol. 2013; 8:1979–1987. [PubMed: 23886564]

142. Frishberg Y, et al. Dent’s disease manifesting as focal glomerulosclerosis: is it the tip of the iceberg? Pediatr. Nephrol. 2009; 24:2369–2373. [PubMed: 19806368]

143. Demoulin N, et al. Gitelman syndrome and glomerular proteinuria: a link between loss of sodium-chloride cotransporter and podocyte dysfunction? Nephrol. Dial. Transplant. 2014; 29(Suppl. 4):iv117–iv120. [PubMed: 25165177]

144. Hanevold C, Mian A, Dalton R. C1q nephropathy in association with Gitelman syndrome: a case report. Pediatr. Nephrol. 2006; 21:1904–1908. [PubMed: 16955279]

145. Blanchard A, et al. Gitelman syndrome: consensus and guidance from a Kidney Disease: Improving Global Outcomes (KDIGO) Controversies Conference. Kidney Int. 2017; 91:24–33. [PubMed: 28003083]

146. Lata, S., et al. Whole exome sequencing in adults with chronic kidney disease: a pilot study. Ann. Intern. Med. 2017. http://dx.doi.org/10.7326/M17-1319

147. Soden SE, et al. Effectiveness of exome and genome sequencing guided by acuity of illness for diagnosis of neurodevelopmental disorders. Sci. Transl Med. 2014; 6:265ra168.

148. Choi M, et al. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proc. Natl Acad. Sci. USA. 2009; 106:19096–19101. This article is among the first papers published showing the utility of WES to achieve a specific diagnosis for patients presenting with nondiagnostic phenotypes. [PubMed: 19861545]

149. Wuttke M, et al. A COL4A5 mutation with glomerular disease and signs of chronic thrombotic microangiopathy. Clin. Kidney J. 2015; 8:690–694. [PubMed: 26613025]

150. Nakata T, et al. Steroid-resistant nephrotic syndrome as the initial presentation of nail-patella syndrome: a case of a *de novo* LMX1B mutation. BMC Nephrol. 2017; 18:100. [PubMed: 28335748]

151. Isnard P, et al. Karyomegalic interstitial nephritis: a case report and review of the literature. Medicine. 2016; 95:e3349. [PubMed: 27196444]

152. Vivante A, et al. Exome sequencing discerns syndromes in patients from consanguineous families with congenital anomalies of the kidneys and urinary tract. J. Am. Soc. Nephrol. 2017; 28:69–75. This article describes the successful use of WES to achieve a molecular diagnosis in genetically unresolved cases of CAKUT. Notably, 44% of diagnosed patients had mutations in genes not classically associated with CAKUT, highlighting the utility of genome-wide testing to pinpoint a specific — and often unexpected — disease aetiology among patients with nonspecific phenotypes. [PubMed: 27151922]

153. Bockenhauer D, Bichet DG. Urinary concentration: different ways to open and close the tap. Pediatr. Nephrol. 2014; 29:1297–1303. [PubMed: 23736674]

154. Emma F, et al. Nephropathic cystinosis: an international consensus document. Nephrol. Dial. Transplant. 2014; 29(Suppl. 4):iv87–iv94. [PubMed: 25165189]

155. Cochat P, Rumsby G. Primary hyperoxaluria. N. Engl. J. Med. 2013; 369:649–658. [PubMed: 23944302]

156. Aymé S, et al. Common elements in rare kidney diseases: conclusions from a kidney disease: Improving Global Outcomes (KDIGO) Controversies Conference. Kidney Int. 2017; 92:796–808. [PubMed: 28938953]

157. Savige J, et al. Alport syndrome in women and girls. Clin. J. Am. Soc. Nephrol. 2016; 11:1713–1720. [PubMed: 27287265]

158. Wang RY, Lelis A, Mirocha J, Wilcox WR. Heterozygous Fabry women are not just carriers, but have a significant burden of disease and impaired quality of life. Genet. Med. 2007; 9:34–45. [PubMed: 17224688]

159. Terryn W, et al. Fabry nephropathy: indications for screening and guidance for diagnosis and treatment by the European Renal Best Practice. Nephrol. Dial. Transplant. 2013; 28:505–517. [PubMed: 23234755]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

160. Lentine KL, et al. KDIGO clinical practice guideline on the evaluation and care of living kidney donors. Transplantation. 2017; 101:S1–S109.

161. Cornec-Le Gall E, et al. Type of PKD1 mutation influences renal outcome in ADPKD. J. Am. Soc. Nephrol. 2013; 24:1006–1013. [PubMed: 23431072]

162. Hwang YH, et al. Refining genotype–phenotype correlation in autosomal dominant polycystic kidney disease. J. Am. Soc. Nephrol. 2016; 27:1861–1868. [PubMed: 26453610]

163. Gunay-Aygun M, et al. Hepatorenal findings in obligate heterozygotes for autosomal recessive polycystic kidney disease. Mol. Genet. Metab. 2011; 104:677–681. [PubMed: 21945273]

164. Zhang J, et al. Incomplete distal renal tubular acidosis from a heterozygous mutation of the V-ATPase B1 subunit. Am. J. Physiol. Renal Physiol. 2014; 307:F1063–F1071. [PubMed: 25164082]

165. Edwards N, et al. A novel LMX1B mutation in a family with end-stage renal disease of ‘unknown cause’. Clin. Kidney J. 2015; 8:113–119. [PubMed: 25713721]

166. Munch J, Grohmann M, Lindner TH, Bergmann C, Halbritter J. Diagnosing FSGS without kidney biopsy — a novel INF2-mutation in a family with ESRD of unknown origin. BMC Med. Genet. 2016; 17:73. [PubMed: 27733133]

167. Ellingford JM, et al. Pinpointing clinical diagnosis through whole exome sequencing to direct patient care: a case of Senior-Loken syndrome. Lancet. 2015; 385:1916. This case report illustrates the potential of genetic testing for early diagnosis in a patient with undetected CKD, enabling preplanned initiation of dialysis, appropriate donor selection for renal transplantation, and surveillance of family members at risk. [PubMed: 25987160]

168. Quaglia M, et al. Unexpectedly high prevalence of rare genetic disorders in kidney transplant recipients with an unknown causal nephropathy. Clin. Transplant. 2014; 28:995–1003. This retrospective study of 911 renal transplant recipients highlights the fact that many patients who progress to ESRD have no clear clinical diagnosis of their renal disease and that rare monogenic forms of nephropathy may be found in a notable fraction of these patients. These findings highlight the potential utility of genetic diagnostics for cases of CKD of unknown aetiology. [PubMed: 24961278]

169. Berg JS. Genome-scale sequencing in clinical care: establishing molecular diagnoses and measuring value. JAMA. 2014; 312:1865–1867. [PubMed: 25326641]

170. Lazaridis KN, et al. Outcome of whole exome sequencing for diagnostic odyssey cases of an individualized medicine clinic: the Mayo Clinic Experience. Mayo Clin. Proc. 2016; 91:297–307. [PubMed: 26944241]

171. ACMG Board of Directors. Clinical utility of genetic and genomic services: a position statement of the American College of Medical Genetics and Genomics. Genet. Med. 2015; 17:505–507. [PubMed: 25764213]

172. Parker S. The pooling of manpower and resources through the establishment of European reference networks and rare disease patient registries is a necessary area of collaboration for rare renal disorders. Nephrol. Dial. Transplant. 2014; 29(Suppl. 4):iv9–iv14. [PubMed: 25165190]

173. European Reference Networks. ERKNet: The European Rare Kidney Disease Reference Network. ERKNet. 2017. https://www.erknet.org/index.php?id=home

174. Krischer JP, Gopal-Srivastava R, Groft SC, Eckstein DJ. Rare Diseases Clinical Research Network. The Rare Diseases Clinical Research Network’s organization and approach to observational research and health outcomes research. J. Gen. Intern. Med. 2014; 29(Suppl. 3):S739–S744. [PubMed: 25029976]

175. UK Kidney Research Consortium. Establishing an infrastructure to support the development and delivery of clinical research in patients with kidney disease. Clin. Med. (Lond.). 2015; 15:415–419. [PubMed: 26430177]

176. US National Library of Medicine. ClinicalTrials.gov. 2016. https://clinicaltrials.gov/ct2/show/NCT02855268

177. US National Library of Medicine. ClinicalTrials.gov. 2016. https://clinicaltrials.gov/ct2/show/NCT02795325

178. Alport Syndrome Foundation. AlportSyndrome.org. 2017. http://alportsyndrome.org/

179. PKD Foundation. PKDCure.org. 2017. https://pkdcure.org/

180. UKD Foundation. UKDCure.org. 2017. http://www.ukdcure.org/

181. Hodson EM, Wong SC, Willis NS, Craig JC. Interventions for idiopathic steroid-resistant nephrotic syndrome in children. Cochrane Database Syst. Rev. 2016; 10:CD003594. [PubMed: 27726125]

182. Richards S, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet. Med. 2015; 17:405–424. In this article, the American College of Medical Genetics puts forth guidelines for diagnostic sequence interpretation. The lines of evidence and variant classification scheme discussed offer a helpful framework to assess variant pathogenicity; the authors also discuss the importance of integrating a patient’s genetic and clinical data to achieve a genetic diagnosis. [PubMed: 25741868]

183. MacArthur DG, et al. Guidelines for investigating causality of sequence variants in human disease. Nature. 2014; 508:469–476. This article provides general standards for sequence interpretation at both the gene level and variant level and emphasizes the importance of using quantitative methods to enable more accurate and reproducible genetic assessment. [PubMed: 24759409]

184. Strande NT, Berg JS. Defining the clinical value of a genomic diagnosis in the era of next-generation sequencing. Annu. Rev. Genomics Hum. Genet. 2016; 17:303–332. [PubMed: 27362341]

185. Parsa A, et al. Common variants in Mendelian kidney disease genes and their association with renal function. J. Am. Soc. Nephrol. 2013; 24:2105–2117. [PubMed: 24029420]

186. Clemente M, et al. Hyperinsulinaemic hypoglycaemia, renal Fanconi syndrome and liver disease due to a mutation in the HNF4A gene. Endocrinol. Diabetes Metab. Case Rep. 2017; 2017:16–0133.

187. Hamilton AJ, et al. The HNF4A R76W mutation causes atypical dominant Fanconi syndrome in addition to a beta cell phenotype. J. Med. Genet. 2014; 51:165–169. [PubMed: 24285859]

188. Bekheirnia MR, et al. Whole-exome sequencing in the molecular diagnosis of individuals with congenital anomalies of the kidney and urinary tract and identification of a new causative gene. Genet. Med. 2016; 19:412–420. [PubMed: 27657687]

189. Jiang S, et al. Lack of major involvement of human uroplakin genes in vesicoureteral reflux: implications for disease heterogeneity. Kidney Int. 2004; 66:10–19. [PubMed: 15200408]

190. Kelly H, et al. Uroplakin III is not a major candidate gene for primary vesicoureteral reflux. Eur. J. Hum. Genet. 2005; 13:500–502. [PubMed: 15523493]

191. Hennekam RC, Biesecker LG. Next-generation sequencing demands next-generation phenotyping. Hum. Mutat. 2012; 33:884–886. [PubMed: 22457028]

192. Shashi V, et al. Practical considerations in the clinical application of whole-exome sequencing. Clin. Genet. 2016; 89:173–181. [PubMed: 25678066]

193. Lek M, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016; 536:285–291. [PubMed: 27535533]

194. Cassa CA, Tong MY, Jordan DM. Large numbers of genetic variants considered to be pathogenic are common in asymptomatic individuals. Hum. Mutat. 2013; 34:1216–1220. [PubMed: 23818451]

195. Piton A, Redin C, Mandel JL. XLID-causing mutations and associated genes challenged in light of data from large-scale human exome sequencing. Am. J. Hum. Genet. 2013; 93:368–383. [PubMed: 23871722]

196. Shearer AE, et al. Utilizing ethnic-specific differences in minor allele frequency to recategorize reported pathogenic deafness variants. Am. J. Hum. Genet. 2014; 95:445–453. [PubMed: 25262649]

197. Abdelhak S, et al. Clustering of mutations responsible for branchio-oto-renal (BOR) syndrome in the eyes absent homologous region (eyaHR) of EYA1. Hum. Mol. Genet. 1997; 6:2247–2255. [PubMed: 9361030]

198. Chang EH, et al. Branchio-oto-renal syndrome: the mutation spectrum in EYA1 and its phenotypic consequences. Hum. Mutat. 2004; 23:582–589. [PubMed: 15146463]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

199. NM_000503.5(EYA1):c.1460C>T (p.Ser487Leu). Clin Var. 2017. https://www.ncbi.nlm.nih.gov/clinvar/variation/228678/#clinical-assertions
200. Smith, RJH. Branchiootorenal spectrum disorders. Gene Reviews. 2015. https://www.ncbi.nlm.nih.gov/books/NBK1380
201. Amendola LM, et al. Performance of ACMG-AMP variant-interpretation guidelines among nine laboratories in the Clinical Sequencing Exploratory Research Consortium. Am. J. Hum. Genet. 2016; 98:1067–1076. [PubMed: 27181684]
202. Yang S, et al. Sources of discordance among germ-line variant classifications in ClinVar. Genet. Med. 2017; 19:1118–1126. [PubMed: 28569743]
203. Safarova MS, et al. Variability in assigning pathogenicity to incidental findings: insights from LDLR sequence linked to the electronic health record in 1013 individuals. Eur. J. Hum. Genet. 2017; 25:410–415. [PubMed: 28145427]
204. Harrison SM, et al. Clinical laboratories collaborate to resolve differences in variant interpretations submitted to ClinVar. Genet. Med. 2017; 19:1096–1104. [PubMed: 28301460]
205. Strande NT, et al. Evaluating the clinical validity of gene–disease associations: an evidence-based framework developed by the Clinical Genome Resource. Am. J. Hum. Genet. 2017; 100:895–906. [PubMed: 28552198]
206. Patel RY, et al. ClinGen Pathogenicity Calculator: a configurable system for assessing pathogenicity of genetic variants. Genome Med. 2017; 9:3. [PubMed: 28081714]
207. Li Q, Wang K. InterVar: clinical interpretation of genetic variants by the 2015 ACMG-AMP guidelines. Am. J. Hum. Genet. 2017; 100:267–280. [PubMed: 28132688]
208. Knoppers BM, Zawati MH, Senecal K. Return of genetic testing results in the era of whole-genome sequencing. Nat. Rev. Genet. 2015; 16:553–559. [PubMed: 26239711]
209. Demmer LA, Waggoner DJ. Professional medical education and genomics. Annu. Rev. Genomics Hum. Genet. 2014; 15:507–516. [PubMed: 24635717]
210. Meagher KM, McGowan ML, Settersten RA, Fishman JR, Juengst ET. Precisely where are we going? Charting the new terrain of precision prevention. Annu. Rev. Genomics Hum. Genet. 2017; 18:369–387. This review examines central ELSIs of expanding genetic testing to a population-wide scale and highlights the need for study of both genetic and environmental determinants of health. [PubMed: 28441061]
211. Berg JS, Khoury MJ, Evans JP. Deploying whole genome sequencing in clinical practice and public health: meeting the challenge one bin at a time. Genet. Med. 2011; 13:499–504. [PubMed: 21558861]
212. Burke W. Genetic tests: clinical validity and clinical utility. Curr. Protoc. Hum. Genet. 2014; 81:9.15.1–9.15.8.
213. Green RC, et al. ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. Genet. Med. 2013; 15:565–574. [PubMed: 23788249]
214. Kalia SS, et al. Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics. Genet. Med. 2017; 19:249–255. [PubMed: 27854360]
215. van El CG, et al. Whole-genome sequencing in health care: recommendations of the European Society of Human Genetics. Eur. J. Hum. Genet. 2013; 21:580–584. [PubMed: 23676617]
216. Amendola LM, et al. Actionable exomic incidental findings in 6503 participants: challenges of variant classification. Genome Res. 2015; 25:305–315. [PubMed: 25637381]
217. Olfson E, et al. Identification of medically actionable secondary findings in the 1000 Genomes. PLoS ONE. 2015; 10:e0135193. [PubMed: 26332594]
218. Birdwell KA, et al. Clinical Pharmacogenetics Implementation Consortium (CPIC) guidelines for CYP3A5 genotype and tacrolimus dosing. Clin. Pharmacol. Ther. 2015; 98:19–24. [PubMed: 25801146]
219. Relling MV, et al. Clinical Pharmacogenetics Implementation Consortium guidelines for thiopurine methyltransferase genotype and thiopurine dosing. Clin. Pharmacol. Ther. 2011; 89:387–391. [PubMed: 21270794]
220. Hunter JE, et al. A standardized, evidence-based protocol to assess clinical actionability of genetic disorders associated with genomic variation. Genet. Med. 2016; 18:1258–1268. This

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

article gives a systematic workflow to assess the medical actionability of a given genetic disease, which can be used to help determine which results may merit return as medically actionable secondary findings. [PubMed: 27124788]

221. O'Daniel JM, et al. A survey of current practices for genomic sequencing test interpretation and reporting processes in US laboratories. Genet. Med. 2017; 19:575–582. [PubMed: 27811861]
222. Otten E, et al. Is there a duty to recontact in light of new genetic technologies? A systematic review of the literature. Genet. Med. 2015; 17:668–678. [PubMed: 25503495]
223. Pyeritz RE. The coming explosion in genetic testing — is there a duty to recontact? N. Engl. J. Med. 2011; 365:1367–1369. [PubMed: 21995382]
224. ACMG Board of Directors. Points to consider in the clinical application of genomic sequencing. Genet. Med. 2012; 14:759–761. [PubMed: 22863877]
225. Shahmirzadi L, et al. Patient decisions for disclosure of secondary findings among the first 200 individuals undergoing clinical diagnostic exome sequencing. Genet. Med. 2014; 16:395–399. [PubMed: 24113345]
226. Delikurt T, Williamson GR, Anastasiadou V, Skirton H. A systematic review of factors that act as barriers to patient referral to genetic services. Eur. J. Hum. Genet. 2015; 23:739–745. [PubMed: 25205405]
227. Cichon M, Feldman GL. Opportunities to improve recruitment into medical genetics residency programs: survey results of program directors and medical genetics residents. Genet. Med. 2014; 16:413–418. [PubMed: 24136619]
228. Plunkett-Rondeau J, Hyland K, Dasgupta S. Training future physicians in the era of genomic medicine: trends in undergraduate medical genetics education. Genet. Med. 2015; 17:927–934. [PubMed: 25674779]
229. Berns JS. A survey-based evaluation of self-perceived competency after nephrology fellowship training. Clin. J. Am. Soc. Nephrol. 2010; 5:490–496. [PubMed: 20089491]
230. Murray MF. Educating physicians in the era of genomic medicine. Genome Med. 2014; 6:45. [PubMed: 25031626]
231. Kentwell M, et al. Mainstreaming cancer genetics: a model integrating germline BRCA testing into routine ovarian cancer clinics. Gynecol. Oncol. 2017; 145:130–136. [PubMed: 28162234]
232. Rhodes A, et al. Minding the genes: a multidisciplinary approach towards genetic assessment of cardiovascular disease. J. Genet. Couns. 2017; 26:224–231. [PubMed: 27695998]
233. Green RC, et al. Clinical Sequencing Exploratory Research Consortium: accelerating evidence-based practice of genomic medicine. Am. J. Hum. Genet. 2016; 98:1051–1066. [PubMed: 27181682]
234. Skirton H, et al. Genetic education and the challenge of genomic medicine: development of core competences to support preparation of health professionals in Europe. Eur. J. Hum. Genet. 2010; 18:972–977. [PubMed: 20442748]
235. Gross O, et al. Safety and efficacy of the ACE-inhibitor ramipril in alport syndrome: the double-blind, randomized, placebo-controlled, multicenter phase III EARLY PRO-TECT Alport trial in pediatric patients. ISRN Pediatr. 2012; 2012:436046. [PubMed: 22811928]
236. US National Library of Medicine. ClinicalTrials.gov. 2017. https://clinicaltrials.gov/ct2/show/NCT01485978
237. Dumitrescu L, et al. Genome-wide study of resistant hypertension identified from electronic health records. PLoS ONE. 2017; 12:e0171745. [PubMed: 28222112]
238. Facio FM, et al. Motivators for participation in a whole-genome sequencing study: implications for translational genomics research. Eur. J. Hum. Genet. 2011; 19:1213–1217. [PubMed: 21731059]
239. Hallowell N, et al. An investigation of patients' motivations for their participation in genetics-related research. J. Med. Ethics. 2010; 36:37–45. [PubMed: 20026692]
240. US Department of Health and Human Services. How does genetic testing in a research setting differ from clinical genetic testing?. Genetics Home Reference. 2017. https://ghr.nlm.nih.gov/primer/testing/researchtesting

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

241. Ferreira-Gonzalez A, et al. US system of oversight for genetic testing: a report from the Secretary's Advisory Committee on Genetics, Health and Society. Per. Med. 2008; 5:521–528. [PubMed: 20490368]

242. Jarvik GP, et al. Return of genomic results to research participants: the floor, the ceiling, and the choices in between. Am. J. Hum. Genet. 2014; 94:818–826. This article discusses the practical and ethical questions surrounding the return of medically actionable results to participants in genomic research studies and gives consensus guidelines to help navigate this complex process. [PubMed: 24814192]

243. Aronson SJ, Rehm HL. Building the foundation for genomics in precision medicine. Nature. 2015; 526:336–342. [PubMed: 26469044]

244. Lyon GJ. Personalized medicine: bring clinical standards to human-genetics research. Nature. 2012; 482:300–301. [PubMed: 22337032]

245. Crews DC, Liu Y, Boulware LE. Disparities in the burden, outcomes, and care of chronic kidney disease. Curr. Opin. Nephrol. Hypertens. 2014; 23:298–305. [PubMed: 24662984]

246. Patzer RE, McClellan WM. Influence of race, ethnicity and socioeconomic status on kidney disease. Nat. Rev. Nephrol. 2012; 8:533–541. [PubMed: 22735764]

247. Webb BD, et al. A founder mutation in COL4A3 causes autosomal recessive Alport syndrome in the Ashkenazi Jewish population. Clin. Genet. 2014; 86:155–160. [PubMed: 23927549]

248. Verlander PC, et al. Carrier frequency of the IVS4 + 4 A-->T mutation of the Fanconi anemia gene FAC in the Ashkenazi Jewish population. Blood. 1995; 86:4034–4038. [PubMed: 7492758]

249. Fedick A, Jalas C, Treff NR. A deleterious mutation in the PEX2 gene causes Zellweger syndrome in individuals of Ashkenazi Jewish descent. Clin. Genet. 2014; 85:343–346. [PubMed: 23590336]

250. Friedman DJ, Pollak MR. Apolipoprotein L1 and kidney disease in African Americans. Trends Endocrinol. Metab. 2016; 27:204–215. [PubMed: 26947522]

251. Genovese G, et al. Association of trypanolytic ApoL1 variants with kidney disease in African Americans. Science. 2010; 329:841–845. [PubMed: 20647424]

252. Kruzel-Davila E, Wasser WG, Aviram S, Skorecki K. APOL1 nephropathy: from gene to mechanisms of kidney injury. Nephrol. Dial. Transplant. 2016; 31:349–358. [PubMed: 25561578]

253. Kramer HJ, et al. African ancestry-specific alleles and kidney disease risk in Hispanics/Latinos. J. Am. Soc. Nephrol. 2017; 28:915–922. This article demonstrates that the *APOL1* risk alleles and/or sickle cell trait are associated with increased CKD risk amongst self-reported Hispanic individuals, but their prevalence varies widely between the Caribbean and non-Caribbean subpopulations owing to differences in genetic ancestry. These findings highlight the importance of detailed consideration of genetic ancestry when assessing the impact of sequence variants on disease risk within a population. [PubMed: 27650483]

254. Naik RP, et al. Association of sickle cell trait with chronic kidney disease and albuminuria in African Americans. JAMA. 2014; 312:2115–2125. [PubMed: 25393378]

255. Petrovski S, Goldstein DB. Unequal representation of genetic variation across ancestry groups creates healthcare inequality in the application of precision medicine. Genome Biol. 2016; 17:157. [PubMed: 27418169]

256. Manrai AK, et al. Genetic misdiagnoses and the potential for health disparities. N. Engl. J. Med. 2016; 375:655–665. This article describes genetic misdiagnoses amongst African-American patients with suspected familial hypertrophic cardiomyopathy, highlighting the importance of interpreting sequence data in the context of patients' genetic ancestry. [PubMed: 27532831]

257. Need AC, Goldstein DB. Next generation disparities in human genomics: concerns and remedies. Trends Genet. 2009; 25:489–494. [PubMed: 19836853]

258. Popejoy AB, Fullerton SM. Genomics is failing on diversity. Nature. 2016; 538:161–164. [PubMed: 27734877]

259. Shea L, Newschaffer CJ, Xie M, Myers SM, Mandell DS. Genetic testing and genetic counseling among Medicaid-enrolled children with autism spectrum disorder in 2001 and 2007. Hum. Genet. 2014; 133:111–116. [PubMed: 24036677]

260. Hann KEJ, et al. Awareness, knowledge, perceptions, and attitudes towards genetic testing for cancer risk among ethnic minority groups: a systematic review. BMC Public Health. 2017; 17:503. [PubMed: 28545429]

261. Halbert CH, McDonald JA, Magwood G, Jefferson M. Beliefs about genetically targeted care in African Americans. J. Natl Med. Assoc. 2017; 109:98–106. [PubMed: 28599763]

262. National Institutes of Health. NIH funds precision medicine research with a focus on health disparities. NIH. 2016. https://www.nih.gov/news-events/news-releases/nih-funds-precision-medicine-research-focus-health-disparities

263. Kingsmore SF, et al. Next-generation community genetics for low- and middle-income countries. Genome Med. 2012; 4:25. [PubMed: 22458566]

264. Tekola-Ayele F, Rotimi CN. Translational genomics in low- and middle-income countries: opportunities and challenges. Public Health Genomics. 2015; 18:242–247. [PubMed: 26138992]

265. World Health Organization. Community genetics services: report of a WHO consultation on community genetics in low- and middle-income countries. WHO; 2010.

266. Obrador GT, et al. Genetic and environmental risk factors for chronic kidney disease. Kidney Int. Suppl. 2017; 7:88–106.

267. Maltese PE, et al. Genetic tests for low- and middle-income countries: a literature review. Genet. Mol. Res. 2017; 16 gmr16019466.

268. Bogershausen N, et al. An unusual presentation of Kabuki syndrome with orbital cysts, microphthalmia, and cholestasis with bile duct paucity. Am. J. Med. Genet. A. 2016; 170A:3282–3288.

269. Moosa S, et al. Novel IFT122 mutations in three Argentinian patients with cranioectodermal dysplasia: expanding the mutational spectrum. Am. J. Med. Genet. A. 2016; 170A:1295–1301. [PubMed: 26792575]

270. Osafo C, et al. Human Heredity and Health (H3) in Africa Kidney Disease Research Network: a focus on methods in sub-Saharan Africa. Clin. J. Am. Soc. Nephrol. 2015; 10:2279–2287. [PubMed: 26138261]

271. Illumina. H3Africa Consortium Array Available Soon. Illumina. 2016. https://www.illumina.com/company/news-center/feature-articles/h3africa-consortium-array-available-soon-.html

272. Bredenoord AL, de Vries MC, van Delden H. The right to an open future concerning genetic information. Am. J. Bioeth. 2014; 14:21–23.

273. Wilfond BS, Fernandez CV, Green RC. Disclosing secondary findings from pediatric sequencing to families: considering the "benefit to families". J. Law Med. Eth. 2015; 43:552–558.

274. Committee on Bioethics *et al*. Ethical and policy issues in genetic testing and screening of children. Pediatrics. 2013; 131:620–622. [PubMed: 23428972]

275. Ross LF, et al. Technical report: ethical and policy issues in genetic testing and screening of children. Genet. Med. 2013; 15:234–245. [PubMed: 23429433]

276. Hufnagel SB, Martin LJ, Cassedy A, Hopkin RJ, Antommaria AH. Adolescents' preferences regarding disclosure of incidental findings in genomic sequencing that are not medically actionable in childhood. Am. J. Med. Genet. A. 2016; 170A:2083–2088.

277. Otlowski M, Taylor S, Bombard Y. Genetic discrimination: international perspectives. Annu. Rev. Genomics Hum. Genet. 2012; 13:433–454. [PubMed: 22607273]

278. Joly Y, Feze IN, Song L, Knoppers BM. Comparative approaches to genetic discrimination: chasing shadows? Trends Genet. 2017; 33:299–302. [PubMed: 28365141]

279. Yoshizawa G, et al. ELSI practices in genomic research in East Asia: implications for research collaboration and public participation. Genome Med. 2014; 6:39. [PubMed: 24944586]

280. Joly Y, Ngueng Feze I, Simard J. Genetic discrimination and life insurance: a systematic review of the evidence. BMC Med. 2013; 11:25. [PubMed: 23369270]

281. Council of Europe. Recommendation of the Committee of Ministers to the member States on the processing of personal health-related data for insurance purposes, including data resulting from genetic tests. Quotidiano Sanita. 2016. http://www.quotidianosanita.it/allegati/allegato2027308.pdf

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

282. Zlotogora J. Genetics and genomic medicine in Israel. Mol. Genet. Genomics Med. 2014; 2:85–94.

283. Association of British Insurers. Concordat and moratorium on genetics and insurance. ABI. 2014. https://www.abi.org.uk/globalassets/sitecore/files/documents/publications/public/2014/genetics/concordat-and-moratorium-on-genetics-and-insurance.pdf

284. Canadian Life and Health Insurance Association. Industry code on genetics testing information for insurance underwriting. CLHIA. 2017. http://www.clhia.ca/domino/html/clhia/clhia_lp4w_lnd_webstation.nsf/page/E79687482615DFA485257D5D00682400OpenDocument

285. Hudson KL, Pollitz K. Undermining genetic privacy? Employee wellness programs and the law. N. Engl. J. Med. 2017; 377:1–3. [PubMed: 28537794]

286. Manolio TA, et al. Bedside back to bench: building bridges between basic and clinical genomic research. Cell. 2017; 169:6–12. This article highlights key priorities for integrating genomic research with clinical care and proposes collaborative efforts between researchers, clinicians, and patients to help address them. [PubMed: 28340351]

287. Tzur S, et al. Missense mutations in the *APOL1* gene are highly associated with end stage kidney disease risk previously attributed to the *MYH9* gene. Hum. Genet. 2010; 128:345–350. [PubMed: 20635188]

288. McLean NO, Robinson TW, Freedman BI. APOL1 gene kidney risk variants and cardiovascular disease: getting to the heart of the matter. Am. J. Kidney Dis. 2017; 70:281–289. [PubMed: 28143671]

289. Freedman BI, et al. Association of APOL1 variants with mild kidney disease in the first-degree relatives of African American patients with non-diabetic end-stage renal disease. Kidney Int. 2012; 82:805–811. [PubMed: 22695330]

290. Divers J, et al. Gene–gene interactions in APOL1-associated nephropathy. Nephrol. Dial. Transplant. 2014; 29:587–594. [PubMed: 24157943]

291. Kopp JB, et al. APOL1 genetic variants in focal segmental glomerulosclerosis and HIV-associated nephropathy. J. Am. Soc. Nephrol. 2011; 22:2129–2137. [PubMed: 21997394]

292. Kasembeli AN, et al. APOL1 risk variants are strongly associated with HIV-associated nephropathy in black South Africans. J. Am. Soc. Nephrol. 2015; 26:2882–2890. [PubMed: 25788523]

293. Newell KA, et al. Integrating APOL1 gene variants into renal transplantation: considerations arising from the American Society of Transplantation Expert Conference. Am. J. Transplant. 2017; 17:901–911. [PubMed: 27997071]

294. Harewood L, et al. Bilateral renal agenesis/hypoplasia/dysplasia (BRAHD): postmortem analysis of 45 cases with breakpoint mapping of two *de novo* translocations. PLoS ONE. 2010; 5:e12375. [PubMed: 20811621]

295. Mansouri MR, et al. Molecular genetic analysis of a *de novo* balanced translocation t(6;17)(p21.31;q11.2) associated with hypospadias and anorectal malformation. Hum. Genet. 2006; 119:162–168. [PubMed: 16395596]

296. Holmberg C, Jalanko H. Congenital nephrotic syndrome and recurrence of proteinuria after renal transplantation. Pediatr. Nephrol. 2014; 29:2309–2317. [PubMed: 24682440]

297. Patrakka J, et al. Recurrence of nephrotic syndrome in kidney grafts of patients with congenital nephrotic syndrome of the Finnish type: role of nephrin. Transplantation. 2002; 73:394–403. [PubMed: 11884936]

298. Ashraf S, et al. ADCK4 mutations promote steroid-resistant nephrotic syndrome through CoQ10 biosynthesis disruption. J. Clin. Invest. 2013; 123:5179–5189. [PubMed: 24270420]

299. Heeringa SF, et al. COQ6 mutations in human patients produce nephrotic syndrome with sensorineural deafness. J. Clin. Invest. 2011; 121:2013–2024. [PubMed: 21540551]

300. Kashtan, C. Alport syndrome and thin basement membrane nephropathy. Gene Reviews. 2015. https://www.ncbi.nlm.nih.gov/books/NBK1207/

301. Bekheirnia MR, et al. Genotype-phenotype correlation in X-linked Alport syndrome. J. Am. Soc. Nephrol. 2010; 21:876–883. [PubMed: 20378821]

302. Kidney Disease Improving Global Outcomes Glomerulonephritis Work Group. KDIGO clinical practice guideline for glomerulonephritis. Kidney Int. Suppl. 2012; 2:139–274.

303. Mitchel, MW., et al. 17q12 recurrent deletion syndrome. Gene Reviews. 2016. https://www.ncbi.nlm.nih.gov/books/NBK401562

304. Bollee G, et al. Adenine phosphoribosyltransferase deficiency. Clin. J. Am. Soc. Nephrol. 2012; 7:1521–1527. [PubMed: 22700886]

305. Runolfsdottir HL, Palsson R, Agustsdottir IM, Indridason OS, Edvardsson VO. Kidney disease in adenine phosphoribosyltransferase deficiency. Am. J. Kidney Dis. 2016; 67:431–438. [PubMed: 26724837]

306. Kirmani, S., Young, WF. Hereditary paraganglioma-pheochromocytoma syndromes. Gene Reviews. 2014. https://www.ncbi.nlm.nih.gov/books/NBK1548/

307. Rednam SP, et al. Von Hippel-Lindau and hereditary pheochromocytoma/paraganglioma syndromes: clinical features, genetics, and surveillance recommendations in childhood. Clin. Cancer Res. 2017; 23:e68–e75. [PubMed: 28620007]

308. Bali, DS., Chen, YT., Austin, S., Goldstein, JL. Glycogen storage disease type I. Gene Reviews. 2016. https://www.ncbi.nlm.nih.gov/books/NBK1312/

309. Froissart R, et al. Glucose-6-phosphatase deficiency. Orphanet J. Rare Dis. 2011; 6:27. [PubMed: 21599942]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

Key points

- Inherited aetiologies are responsible for ~10% of adult end-stage renal disease and >70% of paediatric hephropathy; sequencing studies of large cohorts will shed further light on genetic contributions across different forms of kidney disease
- In addition to ending the ‘diagnostic odyssey’, a genetic diagnosis can provide a deeper understanding of disease pathogenesis, inform prognosis, and guide clinical management
- Genetic testing is currently recommended for patients with early-onset nephropathy and/or other clinical features consistent with an inherited form of disease as well as for evaluation of living kidney donors
- Development of disease-specific guidelines and use of population genetic data will help to facilitate accurate clinical sequence interpretation; nevertheless, patient-level assessment results in the continued need for expert judgement
- The broadening clinical use of genetic testing in nephrology has raised questions regarding the return of results, physician education, testing across different patient subpopulations and many other practical and ethical issues
- Interdisciplinary research and dialogue will help to address unresolved challenges and inform the creation of best practice guidelines for genomic medicine in nephrology

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Box 1

Genetic testing for *APOL1* risk alleles

A rapidly developing area in nephrology is the issue of genetic testing for *APOL1* risk alleles: two common coding variants that strongly influence the risk of multiple forms of nephropathy among individuals of sub-Saharan African descent. *APOL1*-mediated disease has been reported to follow a recessive model, with individuals with these two risk variants displaying a 7 to 10-fold higher risk of hypertension-associated end-stage renal disease (ESRD) and a 10 to 17-fold higher risk of focal segmental glomerulosclerosis (FSGS)-associated ESRD than those with one or no risk variants[251,287]. Subsequently, the *APOL1* risk genotypes have also been associated with increased risk of other forms of kidney disease, progression to ESRD, and allograft failure[250,252]. They might also impact risk of cardiovascular disease, although the precise nature of this association remains to be determined as both positive and protective effects have been reported[288].

Importantly, penetrance is variable; although the *APOL1* risk genotypes confer substantially higher odds of nephropathy than nonrisk genotypes, only a minority of individuals with *APOL1* risk genotypes develop kidney disease[289]. This finding suggests that *APOL1*-mediated kidney disease follows a two-hit model, whereby a secondary factor is required in addition to the risk genotype for disease development[250,252]. Such secondary factors might include environmental or genetic modifiers, including risk variants at other loci[290], and immunological triggers, such HIV infection[291,292]. Given the complexity of *APOL1*-mediated kidney disease, the clinical utility of the risk genotypes remains unclear.

The 2017 Kidney Disease: Improving Global Outcomes (KDIGO) guidelines recommend *APOL1* testing for candidate kidney donors but acknowledge that the evidence base is moderate, as insufficient evidence exists regarding the impact of donation on lifetime risk of nephropathy[160]. Additional investigation, including of the long-term outcomes of individuals with *APOL1* risk genotypes, allograft survival from donors with the *APOL1* risk genotypes, and the pathobiology underlying *APOL1*-mediated disease, will help guide how to best utilize *APOL1* risk variant status in nephrology[293].

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

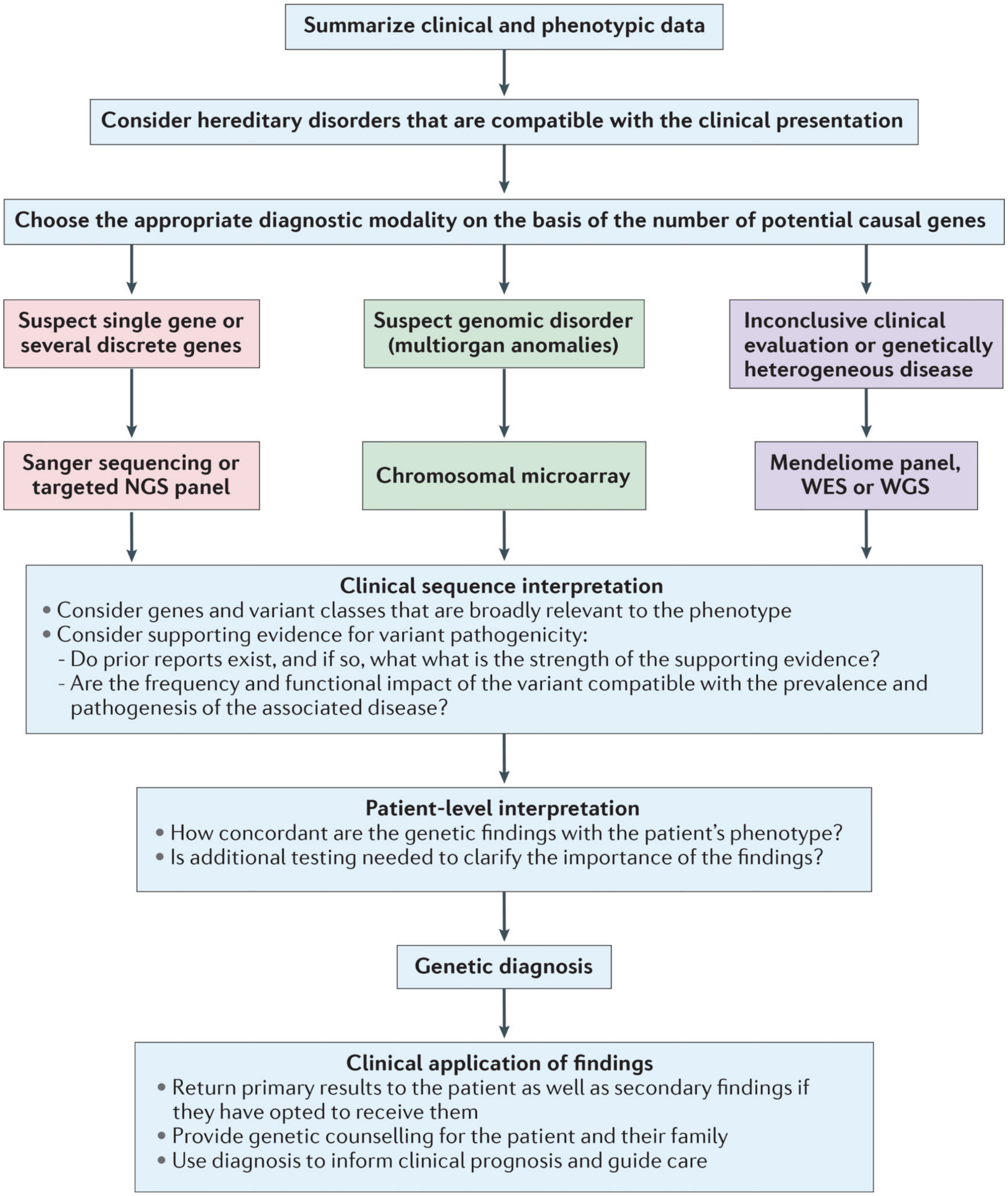

Summarize clinical and phenotypic data
Consider hereditary disorders that are compatible with the clinical presentation
Choose the appropriate diagnostic modality on the basis of the number of potential causal genes
Suspect single gene or several discrete genes
Suspect genomic disorder (multiorgan anomalies)
Inconclusive clinical evaluation or genetically heterogeneous disease
Sanger sequencing or targeted NGS panel
Chromosomal microarray
Mendeliome panel, WES or WGS
Clinical sequence interpretation
• Consider genes and variant classes that are broadly relevant to the phenotype
• Consider supporting evidence for variant pathogenicity:
- Do prior reports exist, and if so, what what is the strength of the supporting evidence?
- Are the frequency and functional impact of the variant compatible with the prevalence and pathogenesis of the associated disease?
Patient-level interpretation
• How concordant are the genetic findings with the patient's phenotype?
• Is additional testing needed to clarify the importance of the findings?
Genetic diagnosis
Clinical application of findings
• Return primary results to the patient as well as secondary findings if they have opted to receive them
• Provide genetic counselling for the patient and their family
• Use diagnosis to inform clinical prognosis and guide care

Figure 1. The genomic nephrology workflow: genetic diagnosis and clinical application

The first step in obtaining a genetic diagnosis for a patient with kidney disease is to characterize their disease phenotype by summarizing their clinical history and other relevant data (for example, findings from biochemical, imaging, and histopathological studies). This phenotype is then used to guide the choice of genetic testing modality. Among patients with genetically heterogeneous disease aetiologies, clinically ambiguous phenotypes, or null results obtained using targeted forms of genetic testing such as Sanger sequencing or targeted next-generation sequencing (NGS) panels, increasingly broad sequencing approaches can be applied, including Mendeliome panels, which can detect variants in all

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

known disease-causing genes; whole-exome sequencing (WES), which can detect variants in all coding regions; and whole-genome sequencing (WGS), which can detect variants in all coding and non-coding regions. Clinical sequence interpretation should be performed according to consensus guidelines[66,67,182]. This process involves identifying genes that are relevant to the phenotype of the patient, prioritizing variants on the basis of prior reports in disease cases as well as compatibility with the prevalence and genetic pathogenesis of the associated disease, and assessing the concordance between the genetic findings and the clinical phenotype. If deemed diagnostic, these primary findings, together with secondary findings if the patient has opted to receive them, can be returned and used to inform prognosis and guide personalized care, including targeted work-up and surveillance, choice of therapy, referral for clinical trials and family counselling.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 1

Major genetic testing modalities: indications and limitations

Modality	Scope	Indications for use	Examples	Advantages	Disadvantages
Sanger sequencing	Detection of SNVs and small indels (<10 bp) within a DNA segment of <1 kb	• Confirmation of NGS findings • Regions refractory to NGS, such as GC-rich, highly repetitive segments • Patients whose phenotype is indicative of a disorder caused by mutations in one specific gene	• Confirm frameshift *COL4A3* variant detected by NGS • Diagnostic testing for Fabry disease • Detect *CTNS* mutation (nephropathic cystinosis) in a patient with corneal cystine crystals and Fanconi syndrome	• High analytical accuracy • Easier and faster sequence interpretation compared with multigene testing enables faster turnaround time • No risk of secondary findings	• Resolution <1 kb; cannot detect larger structural variants • Increasingly time- and cost-inefficient with increasing gene length and/or number of genes tested
Chromosomal microarray	Genome-wide detection of CNVs ≥200–400 kb	Patients with phenotypes commonly resulting from genomic imbalances, such as multiple congenital anomalies	• Detect whole-gene deletion of *HNF1B* in a patient with renal hypodysplasia and autism • Detect 22q11.2 deletion (DiGeorge syndrome) in a patient with renal agenesis and neonatal hypocalcaemia	• Higher resolution enables detection of CNVs missed by karyotyping • Genome-wide CNV detection increases diagnostic sensitivity	• Cannot detect SNVs, indels, and small CNVs • Limited ability to detect balanced chromosomal rearrangements, low-grade somatic mosaicism, and CNVs in certain regions (such as pseudogenes and repetitive elements)
Targeted NGS panels	Detection of SNVs and small indels (<1 kb) within genes of interest for the clinically suspected phenotype	• Patients with phenotypes that are fairly specific for a particular disorder • Disorders with low genetic and/or phenotypic heterogeneity	• Testing *AGXT*, *HOGA1*, and *GRHPR* for primary hyperoxaluria in a patient with childhood-onset calcium oxalate urolithiasis • Testing for *COL4A3*, *COL4A4*, and *COL4A5* mutations in a patient with suspected Alport syndrome	• Can be optimized to ensure sufficient coverage of variants in targeted regions • Interrogation of genes that are related to the clinical indication facilitates interpretation and minimizes risk of secondary findings	• Testing a limited number of genes decreases diagnostic sensitivity, especially for genetically and/or phenotypically heterogeneous disorders • Challenges of panel design (gene selection and need for frequent updates) • Minimal capacity for sequence reanalysis
Whole-exome sequencing	Detection of SNVs and small indels (<1 kb)	• Patients with highly genetically heterogeneous or	• NPHP-RC[122,123] • Diagnosis of congenital chloride	• Unbiased approach increases diagnostic sensitivity	• Lower analytical sensitivity and specificity than whole-genome sequencing

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Modality	Scope	Indications for use	Examples	Advantages	Disadvantages
	within coding regions of the genome	nonspecific phenotypes • CKD of unknown aetiology • Patients left undiagnosed by targeted NGS panels	diarrhoea in an unresolved case of presumed Bartter syndrome[148] • Diagnosis of *LMX1B* glomerulopathy in familial ESRD of unknown origin[165]	• Interrogation of the coding regions that are enriched for known disease-causing mutations is a cost-effective approach to genome-wide testing • Genome-wide scope enables sequence reanalysis and discovery of novel genes	owing to limited coverage of certain regions and inability to accurately call certain types of variants (such as indels) • Approach can lead to multiple candidate variants, increasing time required for interpretation and need for follow-up testing • Burden of secondary findings in genes unrelated to the primary indication for testing
Whole-genome sequencing	Detection of SNVs and small indels (<1 kb) within coding and non-coding regions of the genome	• Patients with highly genetically heterogeneous phenotypes • Patients with nonspecific phenotypes • CKD of unknown aetiology • Patients left undiagnosed by all other genetic testing modalities	• Detection of causal intronic variants, for example, in a genetically unresolved case of Gitelman syndrome[106] • Genetic diagnosis of ADPKD (owing to high sequence homology of *PKD1*)[110] • Detection of causal balanced translocations for congenital anomalies[294,295]	Superior diagnostic and analytical sensitivity to whole-exome sequencing owing to its ability to assess SNVs, indels, and CNVs in coding and non-coding regions and more complete per-base coverage	• Difficulty of interpreting non-coding variants • Large amount of data generated results in substantial time and monetary costs, hindering return of results • Burden of secondary findings in genes unrelated to the primary indication for testing • Burden of long-term sequence data storage

ADPKD, autosomal dominant polycystic kidney disease; ADTKD-*MUC1*, autosomal dominant tubulointerstitial kidney disease due to mutations in *MUC1*; *AGXT,* alanine-glyoxylate aminotransferase; CKD, chronic kidney disease; CNV, copy number variant; COL4A, collagen type IV α-chain; *CTNS*, cytinosin, lysosomal cystine transporter; ESRD, end-stage renal disease; GC, guanine-cytosine; *GRHPR*, glyoxylate and hydroxypyruvate reductase; *HOGA1*, 4-hydroxy-2-oxoglutarate aldolase 1; *HNF1B*, HNF1 homeobox B; *LMX1B*, LIM homeobox transcription factor 1β; *MUC1*, mucin 1, cell surface associated; NGS, next-generation sequencing; NPHP-RC, nephronophthisis-related ciliopathy; *PKD1*, polycystin 1, transient receptor potential channel interacting; SNVs, single-nucleotide variants.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 2

Examples of the clinical utility of genetic diagnoses in nephrology

Indication	**Genetic finding**	**Genetic diagnosis**	**Clinical impact**	**Refs**
Steroid-resistant nephrotic syndrome	Homozygous Fin-major mutation in *NPHS1*	Nephrotic syndrome type 1 (OMIM 256300)	Increased risk of post-transplant disease recurrence	296,297
	COQ2 mutation	CoQ_{10} deficiency 1 (OMIM 607426)	CoQ_{10} supplementation can attenuate proteinuria and extrarenal complications such as encephalopathy	298,299
	COL4A3 or COL4A4 missense mutation	Alport syndrome (OMIM 104200; 203780) or TBMD (OMIM 141200)	• Distinguishes between autosomal (*COL4A3* or *COL4A4*) and X-linked (*COL4A5*) inheritance, informing family counselling • Missense mutations are associated with less severe disease and slower progression to ESRD than loss-of-function mutations • Avoid immunosuppression (a commonly used therapy for nephrotic syndrome)	79, 300–302
Cystic renal dysplasia	17q12 deletion	Renal cysts and diabetes syndrome (OMIM 137920)	Multisystem work-up for associated extrarenal complications, including testing for diabetes, exocrine pancreatic insufficiency, hepatic function, neurological anomalies, and/or neurocognitive impairment	133,134, 303
Nephrolithiasis	*APRT* mutation	APRT deficiency (OMIM 614723)	Xanthine dehydrogenase inhibition to prevent crystalline nephropathy and allograft loss	304,305
Episodic hypertension	*SDHD* mutation	Hereditary paraganglioma-pheochromocytoma syndrome (OMIM 168000)	• Imaging studies to screen for additional tumours • Catecholamine antagonists and/or surgical tumour resection • Knowledge of parent-of-origin effect due to maternal imprinting informs genetic counselling • Lower risk of malignancy than other genetic causes of familial paragangliomas-pheochromocytoma syndromes informs prognosis	306,307
Failure to thrive, hepatomegaly, and hyperuricemia	*G6PC* mutation	Glycogen storage disease Ia (OMIM 232200)	• Dietary therapy (frequent meals, nasogastric tube, and/or raw starch to prevent hypoglycaemia; oral bicarbonate and avoidance of fructose and glucose to prevent acidosis)	308,309

Indication	Genetic finding	Genetic diagnosis	Clinical impact	Refs
			- Surveillance for hepatic adenoma; liver transplant may be needed	

APRT, adenine phosphoribosyltransferase; *COL4A*, collagen type IV α-chain; CoQ$_{10}$, Coenzyme Q$_{10}$; *COQ2*, coenzyme Q2, polyprenyltransferase; ESRD, end-stage renal disease; G6PC, glucose-6-phosphatase catalytic subunit; *NPHS1*, nephrin; SDHD, succinate dehydrogenase complex subunit D; TBMD, thin basement membrane disease.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 3

Resources for diagnostic sequence interpretation and clinical application

Resource	**Description**	**Website(s)**
Summarize patient data		
PhenoTips	• Web interface for clinicians to enter patient data and map symptoms to HPO terminology • Suggests candidate OMIM disorders and associated gene panels on the basis of phenotypic information	https://phenotips.org
Linking Open Data for Rare Diseases	• Web database that helps to code rare diseases by use of Orphanet, HPO, and OMIM data; provides information on epidemiology, inheritance, and characteristic features • Users can browse by disease name, medical specialty, signs or symptoms, and organ system(s) affected	https://bioportal.bioontology.org/projects/LORD
KCCG Patient Archive	• Web interface for clinicians to securely enter and store patients' clinical and genetic data and share these with other clinicians • Stores filtered summary results of genetic testing, facilitating writing of≈clinical genetic notes and return of results	http://www.patientarchive.org
Choose genes to analyse		
OMIM	• Database of >5,000 Mendelian disorders and >15,000 genes derived from curation of the primary literature • Provides an overview of each disorder and the associated genes and selected variants that have previously been reported	https://omim.org
Orphanet	• Database of rare diseases and associated genes, coded using HPO terminology • Diagnostic criteria are included for each disease to help users to gauge the relevance to a patient's phenotype	http://www.orpha.net
PanelApp	• Database of curated gene panels for particular indications, including a variety of renal disorders • Indicates the strength of association of each gene with the relevant phenotype; those with strong evidence are recommended for clinical testing panels	https://panelapp.extge.co.uk/crowdsourcing/PanelApp
ClinGen Gene-Disease Clinical Validity Classification Framework	Framework to assess the strength of association between a gene and a disease on the basis of evidence including the number of reported patients with variants in the gene and supporting experimental data	https://www.clinicalgenome.org/working-groups/gene-curation/projects-initiatives/clinical-validity-classifications
Prioritize variants		
ExAC and gnomAD	• ExAC is a database of WES data from 60,706 individuals; its successor, gnomAD, includes WES data from 123,136 individuals and WGS data from 15,496 individuals, encompassing seven major global populations • ExAC and gnomAD exclude related individuals and/or those with severe early-onset disorders but do include those sequenced as part of various adult-onset disease cohorts; individuals were not phenotyped in detail, so their disease status is unknown	• http://exac.broadinstitute.org • http://gnomad.broadinstitute.org
1000 Genomes	• Database of WGS data from 2,504 individuals from 26 different global populations; the data are of variable quality and include some related individuals • The participants are self-declared as healthy, but additional data regarding their health status was not collected	http://grch37.ensembl.org/index.html
Database of Genomic Variants	Database of >2.5 million structural variants detected in >22,300 healthy individuals; the relatedness status of participants is not mentioned	http://dgv.tcag.ca/dgv/app/home
ClinVar	• Open-access database of inherited and somatic genetic variants detected in individuals across a broad spectrum of disorders; contributors include diagnostic laboratories, other databases and research groups • Each variant is accompanied by assertions regarding its pathogenicity (or lack thereof) for the associated disease phenotype	https://www.ncbi.nlm.nih.gov/clinvar

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Resource	Description	Website(s)
	- The number of stars accompanying an entry reflects the strength of supporting evidence for the classification, which varies considerably	
Human Gene Mutation Database	Database of disease-associated genetic variants, manually curated from review of the scientific literature; access to the most comprehensive version requires a paid subscription	https://www.qiagenbioinformatics.com/products/human-gene-mutation-database
Human Genome Variation Society	Repository for various disease and gene-specific databases, many of which are organized within the LOVD framework; individual databases vary in the amount and quality of information provided regarding the variants deposited	- http://www.hgvs.org/content/databases-tools - http://www.lovd.nl
Variant Effect Predictor	Software tool that predicts the impact of a genetic variant on the encoded protein and assesses its predicted pathogenicity by use of various *in silico* algorithms	http://www.ensembl.org/Tools/VEP
Uniprot	Protein database; users can query genes to see whether a variant is located in a conserved domain or has experimental data supporting an effect on protein function	http://www.uniprot.org
Patient-level interpretation		
MyGene2	- Web portal of phenotypic and genetic data from patients with rare genetic diseases - Users can search by gene or clinical symptom and connect with researchers, clinicians, and patients who share findings in the gene and/or the relevant phenotype	https://www.mygene2.org/MyGene2
Decipher	- Database of sequence and structural variants from patients with rare genetic disorders - Users can search by variant, gene, or disease and directly compare a patient's genetic data to the variants present	https://decipher.sanger.ac.uk
Clinical application of genetic findings		
ClinicalTrials.gov	- Global database of clinical studies, including interventional and observational protocols - Patients, researchers, and clinicians can find opportunities to participate in disease-specific studies and view the results of completed studies	https://clinicaltrials.gov
GeneReviews	- Database that provides clinically relevant and medically actionable information about inherited conditions - Articles are written by domain experts and address aspects including clinical prognosis and management, genetic counselling recommendations, and patient support groups	https://www.ncbi.nlm.nih.gov/books/NBK1116
Orphanet	- Directory of expert centres, physicians, tests, ongoing research and patient organizations and an inventory of orphan drugs - Links users to clinical referral centres, patient registries, and research networks for rare renal diseases	http://www.orpha.net
EURenOmics	International consortium that applies genomic and phenotypic studies of large patient cohorts with functional studies to gain greater insight into rare renal diseases and identify potential therapeutic targets	https://www.eurenomics.eu
American Board of Genetic Counseling	Web portal that enables patients to search for genetic counsellors worldwide by medical specialty and/or location	https://www.abgc.net/about-genetic-counseling/find-a-certified-counselor.aspx

ExAC, Exome Aggregation Consortium; gnomAD, Genome Aggregation Database; HPO, Human Phenotype Ontology; KCCG, Kinghorn Centre for Clinical Genomics; LOVD, Leiden Open Variant Database; OMIM, Online Mendelian Inheritance in Man; WES, whole-exome sequencing; WGS, whole-genome sequencing.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 4

Strategies for bench-to-bedside translation of genetic findings in nephrology

Insight from genetic studies	Clinical need	Knowledge required	Potential strategies to enable translation
A molecular cause can be identified across a variety of clinical presentations	Evidence-based guidelines regarding indications for genetic testing	• Prevalence of genetic forms of nephropathy among patients with kidney disease • Clinical features that predict a genetic form of nephropathy	• Multicentre sequencing studies in large, diverse all-cause kidney disease cohorts • Electronic health record tools to help alert physicians to patients with relevant indications for genetic testing
Expanding the scope of genetic testing increases diagnostic sensitivity but also increases the time and expertise needed for interpretation	Ability to select the appropriate genetic test given the clinical context and provide the phenotypic information needed for sequence interpretation	• The technical and diagnostic advantages and limitations of available genetic testing options • The phenotypic information that should be provided to the diagnostic testing laboratory	• Comprehensive study comparing diagnostic yield and cost-effectiveness between test modalities for different patient populations and clinical indications • Expert working groups to create disease-specific standards for required pretest information, genes assessed, and variant interpretation
In many cases, a one-to-one relationship does not exist between a given genetic mutation and a clinical phenotype	The ability to inform patients and families about the clinical relevance of the genetic variant detected and provide the indicated work-up	• The range of phenotypes associated with mutations in a gene • The factors resulting in genetic pleiotropy and variable expressivity • The prognostic importance of genetic findings	• Comprehensive patient phenotyping, including assessment of extrarenal and renal features • Integration of additional omics data into clinical sequence interpretation • Large cohort studies of patients with genetic forms of kidney disease that investigate the relationship between genotype and long-term clinical outcomes
Genetic findings can substantially impact clinical outcomes, such as health-care utilization, morbidity and mortality, and choice of therapy	Personalized treatment plans and targeted pharmacological agents tailored to specific genetic disorders	• The relationship between a genotype and key clinical outcomes • Disease pathogenesis at a molecular level	• Observational studies assessing the impact of genetic findings on clinical outcomes • Utilization of genetic stratification to better power clinical trials • Application of the biological insights gained from research into development of novel targeted therapeutic agents
The causality and medical actionability of genetic findings are probabilistic and evolving concepts	Appropriate counselling before and after genetic counselling	• The understanding of patients and physicians regarding genetic testing and their preferences for return of results and sequence reanalysis • The types of genetic variants that merit return as medically actionable secondary findings	• Include geneticists and genetic counsellors in multidisciplinary care teams as part of clinical genetic testing • Establish best practice guidelines for sequence reanalysis and patient recontact • Establish a multidisciplinary nephrogenetics working group to create a list of medically actionable secondary renal findings

Jackson, A., Sanders, M., Berry, A., McQuillan, J., Aslett, M., Quail, M.A., Chukualim, B., Capewell, P., MacLeod, A. , Melville, S., Gibson, W., Barry, J.D. , Berriman, M. and Hertz-Fowler, C. (2010) *The genome sequence of Trypanosoma brucei gambiense, causative agent of chronic Human African Trypanosomiasis.* PLoS Neglected Tropical Diseases, 4 (4). e658. ISSN 1935-2727

http://eprints.gla.ac.uk/42670/

Deposited on: 21 December 2011

Enlighten – Research publications by members of the University of Glasgow
http://eprints.gla.ac.uk

The Genome Sequence of *Trypanosoma brucei gambiense*, Causative Agent of Chronic Human African Trypanosomiasis

Andrew P. Jackson[1], Mandy Sanders[1], Andrew Berry[1], Jacqueline McQuillan[1], Martin A. Aslett[1], Michael A. Quail[1], Bridget Chukualim[2], Paul Capewell[3], Annette MacLeod[3], Sara E. Melville[4], Wendy Gibson[5], J. David Barry[3], Matthew Berriman[1], Christiane Hertz-Fowler[1]*

1 Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, United Kingdom, **2** International Trypanotolerance Center, Banjul, The Gambia, **3** Wellcome Centre for Molecular Parasitology, Glasgow Biomedical Research Centre, University of Glasgow, Glasgow, United Kingdom, **4** Hughes Hall, Cambridge, United Kingdom, **5** School of Biological Sciences, University of Bristol, Bristol, United Kingdom

Abstract

Background: *Trypanosoma brucei gambiense* is the causative agent of chronic Human African Trypanosomiasis or sleeping sickness, a disease endemic across often poor and rural areas of Western and Central Africa. We have previously published the genome sequence of a *T. b. brucei* isolate, and have now employed a comparative genomics approach to understand the scale of genomic variation between *T. b. gambiense* and the reference genome. We sought to identify features that were uniquely associated with *T. b. gambiense* and its ability to infect humans.

Methods and Findings: An improved high-quality draft genome sequence for the group 1 *T. b. gambiense* DAL 972 isolate was produced using a whole-genome shotgun strategy. Comparison with *T. b. brucei* showed that sequence identity averages 99.2% in coding regions, and gene order is largely collinear. However, variation associated with segmental duplications and tandem gene arrays suggests some reduction of functional repertoire in *T. b. gambiense* DAL 972. A comparison of the variant surface glycoproteins (VSG) in *T. b. brucei* with all *T. b. gambiense* sequence reads showed that the essential structural repertoire of VSG domains is conserved across *T. brucei*.

Conclusions: This study provides the first estimate of intraspecific genomic variation within *T. brucei*, and so has important consequences for future population genomics studies. We have shown that the *T. b. gambiense* genome corresponds closely with the reference, which should therefore be an effective scaffold for any *T. brucei* genome sequence data. As *VSG* repertoire is also well conserved, it may be feasible to describe the total diversity of variant antigens. While we describe several as yet uncharacterized gene families with predicted cell surface roles that were expanded in number in *T. b. brucei*, no *T. b. gambiense*-specific gene was identified outside of the subtelomeres that could explain the ability to infect humans.

Citation: Jackson AP, Sanders M, Berry A, McQuillan J, Aslett MA, et al. (2010) The Genome Sequence of *Trypanosoma brucei gambiense*, Causative Agent of Chronic Human African Trypanosomiasis. PLoS Negl Trop Dis 4(4): e658. doi:10.1371/journal.pntd.0000658

Editor: Jane M. Carlton, New York University School of Medicine, United States of America

Received October 19, 2009; **Accepted** March 2, 2010; **Published** April 13, 2010

Copyright: © 2010 Jackson et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: The genome sequencing and annotation work was funded by the Wellcome Trust (grant number WT085775/Z/08/Z). The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: chf@sanger.ac.uk

Introduction

Trypanosoma brucei subsp. *gambiense* is the causative agent of Human African Trypanosomiasis (HAT), or sleeping sickness, which is a vector-borne disease restricted to rural areas of sub-Saharan Africa. Trypanosomiasis in humans and livestock imposes substantial morbidity, representing a major impediment of agricultural production in the affected areas [1], and is fatal where untreated. The World Health Organization estimated in 1998 that up to 60 million people are at risk in approximately 250 distinct foci [2], although under-reporting has been estimated as high as 40% in some foci [3]. *T. b. gambiense* is the most clinically relevant sub-species, causing over 90% of all human disease. The gambiense disease is typically chronic, often lasting several years with few severe signs and symptoms until the late stage of nervous system involvement. *T. b. gambiense* is sensitive to treatment with pentamidine (early stage) and eflornithine (late stage), drugs which are frequently ineffective against *T. b. rhodesiense* [4], although the underlying biochemical reasons for these differences are unknown. Combination therapies against the late stage disease have performed encouragingly [5] but few drugs are available. Furthermore, unpleasant and in some cases severe side effects often result in poor patient compliance. Hence, new molecular targets are required to supply current drug discovery programmes [6].

T. brucei is subdivided into three subspecies based on infectivity to humans, pathogenicity and geographical distribution. *T. b. gambiense* and *T. b. rhodesiense* are human pathogens, causing Human African Trypanosomiasis (HAT) in West/Central and East Africa respectively. *T. b. brucei* cannot by definition infect

Author Summary

Sleeping sickness, or Human African Trypanosomiasis, is a disease affecting the health and productivity of poor people in many rural areas of sub-Saharan Africa. The disease is caused by a single-celled flagellate, *Trypanosoma brucei*, which evades the immune system by periodically switching the proteins on its surface. We have produced a genome sequence for *T. brucei gambiense*, which is the particular subspecies causing most disease in humans. We compared this with an existing reference genome for a non-human infecting strain (*T. b. brucei* 927) to identify genes in *T. b. gambiense* that might explain its ability to infect humans and to assess how well the reference performs as a universal plan for all *T. brucei*. The genome sequences differ only due to rare insertions and duplications and homologous genes are over 95% identical on average. The archive of surface antigens that enable the parasite to switch its protein coat is remarkably consistent, even though it evolves very quickly. We identified genes with predicted cell surface functions that are only present in *T. b. brucei* and have evolved rapidly in recent time. These genes might help to explain variation in disease pathology between different *T. brucei* strains in different hosts.

humans and is found in a wide range of wild and domestic mammals. The human pathogens have also been found in various animal species and HAT caused by *T. b. rhodesiense* in East Africa is recognized as a zoonosis. *T. b. gambiense* comprises two groups; a genetically homogeneous group to which the majority of isolates belong (group 1), and a second represented by a handful of isolates from West Africa (group 2). Group 1 *T. b. gambiense* strains have the smallest genomes in the *T. brucei* species complex, having 71–82% of the highest DNA content measured for *T. b. brucei* [7–8]. Pulse-field gel analysis of *T. b. gambiense* chromosomes shows that few if any mini-chromosomes are present compared to the estimated 100 in *T. b. brucei* and *T. b. rhodesiense*, and the mini-chromosomes are also of a smaller size–average 25 kb in *T. b. gambiense* compared to 100 kb in *T. b. brucei* and *T. b. rhodesiense* [8–9].

Perhaps as a consequence of this reduced genome, *T. b. gambiense* also has a restricted repertoire of *Variant Surface Glycoprotein (VSG)* genes [8,10–12]. At any time, bloodstream form trypanosomes possess a surface glycoprotein coat formed through the expression of a single gene from a large archive of *VSGs* [13]. This coat obfuscates the host immune system by shielding the invariant surface epitopes from view and, when an immune response is inevitably raised against the VSG monolayer and the active *VSG* is replaced by another, it allows parasites expressing the novel variant to escape the immune response [13]. This periodic VSG 'switching', or *in situ* activation, is facilitated by transposition of inactive *VSG* into a dedicated expression site at the telomeres by gene conversion [13–15]. Although *VSG* repertoire is clearly very large [16–17], it is not known how *VSG* diversity accumulates over time and between strains. The *SRA* gene encodes a truncated VSG-like protein [18]; it is located within one specific *VSG* expression site and is expressed in Human serum-resistance clones of *T. b. rhodesiense* only [19]. Innate immunity to trypanosomes in Humans is conferred by a trypanolytic factor, apoL1 [20] and *SRA* has acquired a role in neutralizing the toxic effects of this protein [21]. Hence, when transcriptionally activated, *SRA* enables particular *T. b. rhodesiense* clones to infect Humans [22]. *SRA* is absent in *T. b. gambiense* [23] and the underlying basis for the trait of human infectivity here is as yet unknown. In *T. b. gambiense*, as yet the only example of a subspecies-specific gene is *TgsGP*, which encodes a 47 kDa VSG-like receptor protein, and is expressed in the flagellar pocket of bloodstream stage cells [24]. However, *TgsGP* is not associated with human infectivity in *T. b. gambiense* [25].

We produced an improved, high-quality draft genome sequence for *T. b. gambiense* DAL927 with the twin aims of identifying subspecies-specific genomic features that might contribute to our understanding of phenotypic variation and assessing the scale of genomic variation across *T. brucei*. This was achieved through comparison with the *T. b. brucei* 927 reference genome and we sought to evaluate the proficiency of this reference, ahead of the next generation of genome sequencing projects that will compare multiple isolates to scrutinize genetic divergence and genomic rearrangements in relation to disease. Our analyses show that the genome sequence of *T. b. gambiense* corresponds closely in gene order and content to the *T. b. brucei* 927 genome. Intraspecific genomic variation is largely associated with tandem or segmental duplications, among which we identify several subspecies-specific isoforms. Our final objective was to compare the *VSG* repertoires of *T. b. brucei* and *T. b. gambiense*, and so provide the first global perspective of how *VSG* diversity evolves on a genome scale. Details of the genome project describing the 'Minimum Information for Genome Sequences' are available online (http://genomesonline.org/GOLD_CARDS/Gi00917.html).

Methods

Accession numbers

The sequence of the *Trypanosoma brucei gambiense* genome has been submitted to the EMBL database under accession numbers FN554964- FN554974 inclusive.

Trypanosome stocks

The *T. b. gambiense* strain MHOM/CI/86/DAL972 was isolated from a patient in Côte d'Ivoire in 1986 and has been used routinely in laboratory studies since this time [26]. Bloodstream form trypanosomes were fed to tsetse *in vitro* and procyclics from infected midguts were established in culture and subsequently optically cloned. Procyclic form trypanosomes were grown in Cunningham's medium supplemented with 10% v/v heat-inactivated foetal calf serum, 5 µg/ml hemin and 10 µg/ml gentamycin at 27°C. High molecular weight DNA was purified by standard methods of phenol-chloroform extraction and alcohol precipitation.

T. b. gambiense genome sequencing and assembly

T. b. gambiense DNA was randomly sheared, size-selected DNA purified and subcloned into pUC19 plasmids (1.4 kb–4 kb inserts), as well as BAC vectors as previously described [27]. Inserts were sequenced by random sequencing using dye-terminator chemistry on ABI 3730 sequencing machines from both ends to generate paired end reads. There were 369,043 passed paired-end reads, producing roughly eight-fold coverage of the whole genome. Sequence reads were assembled using Phrap (www.phrap.org; P. Green, unpublished). Automated in-house software (Auto-Prefinish) was used to identify primers and clones for additional sequencing to close physical and sequence gaps by oligo-walking. Manual base checking and finishing was carried out using Gap4 (http://www.mrc-lmb.cam.ac.uk/pubseq/manual/gap4_unix_1.html). Regions containing repeat sequences or with an unexpected read depth were manually inspected. The assembled contigs were iteratively ordered and orientated against the *T. brucei* 927 genome sequence, with manual checking. Aided by information from orientated read-pairs, together with additional sequencing from selected large insert clones, we re-examined regions with apparent breaks in chromosomal colinearity for potential assembly errors.

Genome annotation

The human-curated annotation of the *T. b. brucei* 927 reference genome was transferred to the assembled *T. b. gambiense* genome on the basis of BLASTp matches and positional information using custom perl scripts. Subsequently, gene structure and functional annotation were manually inspected and further edited, where appropriate, using the Artemis software [28], as previously detailed [27]. The annotation of the *T. b. gambiense* genome can be viewed and searched via GeneDB (http://www.genedb.org/) and comparative chromosome maps for *T. b. brucei* and *T. b. gambiense* are available at TritrypDB (http://tritrypdb.org; [29]). Chromosomal sequences have been submitted to EMBL with the following accession numbers: FN554964-FN554974 inclusive.

Variation detection from sequence data

The *T. b. gambiense* capillary shotgun reads were aligned against the *T. b. brucei* 927 reference genome using SSAHA2 (http://www.sanger.ac.uk/Software/analysis/SSAHA2/). We discarded reads that mapped to more than one location on the reference genome, as well as pairs of reads that did not map in the correct orientation or to within 20% of the expected insert size of the library. In-house perl scripts were used to identify single nucleotide polymorphisms (SNPs) from the SSAHA alignments that adhered to a modified version of the Neighbourhood Quality Standard (NQS, [30]); we term this AltNQS. According to NQS, an acceptable SNP (or fixed difference) has a phred quality score of ≥23 and the 5 bases on either side of the SNP position have a quality score of ≥15. However, these strict criteria do not allow for multiple mismatches within the 11 bp window. To accommodate the higher levels of polymorphism, our AltNQS adheres to the same rules as NQS but allows for multiple SNPs within the 11 bp alignment window as long as the base quality of each SNP has a phred score of at least 23. To identify regions with significantly high SNP density on each chromosome, non-overlapping windows of 10 kb with at least 50% of read coverage were selected for analysis. For these windows, SNP density was calculated as the number of SNPs divided by the number bases covered in that 10 kb window. Using random sampling we estimated the mean and 97.5% confidence limit of mean SNP density. Regions with a value above the 97.5% quantile were identified as having significantly high SNP density values.

Tandem repeat recombination analysis

Tandem gene arrays in the *T. b. brucei* 927 genome with >3 gene copies have previously been defined, and are known to contain polymorphism that is affected by recombination [31]. We assessed the variation among tandem gene duplicates to identify subspecies-specific genes. For each of these arrays, the coding and 3′ UTR sequences were gathered from the corresponding regions of the *T. b. gambiense* genome sequence. The downstream limit of the 3′ UTR was defined by the polypyrimidine termination motif [32]. All *T. b. brucei* and *T. b. gambiense* sequences were aligned in ClustalX [33] and manually adjusted. Those arrays showing no variation or only corresponding isoforms in both subspecies (i.e., simple orthology) were discarded, leaving just those cases where a disparity in sequence diversity was apparent. To detect any ambiguity in phylogenetic relationships among sequences, each of these alignments was analyzed using SplitsTree v4.3 [34], which applies a Neighbour-Net method [35]) to estimate a phylogenetic network. Genetic distances were corrected for variation in base composition after excluding phylogenetically-uninformative characters. Each alignment was also analyzed using the pair-wise homoplasy index (PHI) test [36] that can detect multiple phylogenetic signals within an alignment and is robust in the presence of rate heterogeneity. A third method, the genetic algorithm for recombination detection (GARD, [37]) was applied to estimate the number and placement of recombination breakpoints along each alignment.

Comparison of the variant surface glycoprotein (VSG) repertoire

1258 predicted VSG protein sequences encoded in the *T. b. brucei* genome were compared with the *T. b. gambiense* 972 read library using pair-wise BLASTp searches. These included 36 *VSG*-related (*VR*) sequences that are structurally distinct from the bulk of canonical *VSG* [17]. Initially, all *VSG*-like sequences were extracted from the *T. b. gambiense* read library using BLASTx against whole VSG protein sequences. Each *T. b. brucei* VSG protein sequence was then individually BLAST-searched against this subset of *VSG*-like reads to determine its closest match in *T. b. gambiense*. A reciprocal comparison was carried out to confirm the relationship. To determine if a given gene was most closely related to a paralog in *T. b. brucei* or to an ortholog in *T. b. gambiense*, each *T. b. brucei* VSG protein sequence was also compared a combined database of *VSG* gene models and *VSG*-like reads using BLASTp. BioLayout Express 3D [38] was used to visualize the relative genetic distances between the 1258 *T. b. brucei* VSG sequences, using the BLAST scores derived from comparisons of each gene with all others, and a 70% cutoff to simplify the resulting network graph. To determine if *VSG* diversity is sub-structured according to life stage, nine VSG sequences known to be associated with metacyclic expression sites were BLAST-searched against all other (bloodstream-expressed) VSG and added to the network.

Results/Discussion

The *T. b. gambiense* genome was whole-genome shotgun sequenced to eight-fold coverage by paired-end capillary sequencing of plasmid and bacterial artificial chromosome (BAC) clones, resulting in an improved high-quality draft sequence. In comparison with the *T. b. brucei* 927 reference sequence, the two genomes are very similar in composition and structure, such that no protein coding sequence unique to *T. b. gambiense* could be found. However, coding sequences unique to *T. b. brucei* were found and the two genomes displayed other subtle differences in the diversity of repetitive regions such as segmental duplications, tandem gene arrays and strand switch regions, which document the scale of genomic variation across *T. brucei* subspecies.

The *T. b. brucei* reference is an effective template for the *T. b. gambiense* genome sequence

The draft genome assembly consists of 1768 contigs larger than 2 kb, amounting to 32.6 Mb of data. Of these, 281 contigs, totaling 22.1 Mb, were ordered and orientated against the *T. b. brucei* 927 reference genome. The remaining contigs encode additional copies of tandemly arrayed gene families as well as genes typically associated with subtelomeres such as expression site associated genes (*ESAGs*), variant surface glycoprotein (*VSG*) genes and the *ingi* transposable element. The gene models and annotation of an initial set of 9898 coding sequences located on core chromosomes (i.e., not in subtelomeres) were transferred to the *T. b. gambiense* genome on the basis of BLASTp matches and positional information using custom perl scripts.

When compared, the *T. b. brucei* and *T. b. gambiense* genome sequences are very similar in terms of content, gene order and sequence identity. The absence of potentially *gambiense*-specific sequences was confirmed by examining a Phrap assembly of those capillary reads that did not map against the *T. brucei* 927 reference genome. Analysis of $\sim$40,000 unmapped sequence reads using BLASTx showed that among them were features homologous to *VSG*, *ESAG* and *RHS* genes, as well as *ingi* retrotransposons, but no additional coding sequences that were missing from *T. b. brucei*. We examined the divergence of coding sequences and a frequency histogram of percentage nucleotide identity (Fig. 1) shows that 86.4% of genes vary by less than 1% from their *T. b. brucei* ortholog (mean average nucleotide identity $= 99.2\%$). Non-coding regions were more divergent, which is unsurprising given that they are probably under weaker purifying selection, but still remained 95.4% identical on average. However, against this general background of correspondence there are 69 pairs of orthologs that display significantly greater evolutionary change, (i.e., they are among the 5% most divergent orthologs with a nucleotide identity $<95.2\%$). 35 of these gene pairs are *VSG* sequences; these surface glycoproteins are exposed to frequent gene conversion and evolve rapidly [16–17], so naturally, they display lower sequence identities of $\sim$60–85%. However, they still display reciprocal top BLAST hits with *T. b. brucei* sequences. Also among these divergent gene pairs are 17 uncharacterized genes, 10 of which are predicted to encode cell-surface targeted proteins. For example, Tb927.5.4010/Tbg972.5.4300 (92.7% identical) and Tb10.70.1280/Tbg972.10.6310 (93.7% identical) are both located at strand-switch regions and encode hypothetical proteins with predicted signal peptides and GPI anchor sites. These genes, which appear to be evolving very quickly, are not found in either *Leishmania major* or *T. cruzi*, indicating that they are specific to African trypanosomes.

A source of variation with potentially important functional consequences is allelic polymorphism. We detected high-confidence SNPs and fixed differences by mapping the *T. b. gambiense* reads to the *T. brucei* 927 reference sequence. Our analysis focused on the non-repetitive component of the genome as firstly, non-identical repeats can appear indistinguishable from SNPs and secondly, repeated regions may be subject to unusual selective pressures (see below). After excluding these sequences, we identified a total of 224,568 putative fixed differences from 19.4 Mb of non-repetitive sequence, i.e. a diversity (π) of 0.0116 nucleotides per site. 92,794 of these differences were in coding regions, 49% of which were non-synonymous. To confirm that the variation identified when mapping the *T. b. gambiense* reads against the *T. b. brucei* 927 were not in fact false-positives due to heterozygosity within the *T. b. brucei* 927 reference sequence itself we also used the available capillary read data from the *T. b. brucei* 927 genome project to identify polymorphism within the published ''haploid'' consensus. Unfortunately, this was only possible for the four chromosomes (1, 9–11) that were originally produced by shotgun sequencing, (rather than a clone walking strategy), since these contain data from two homologous chromosomes at a given locus. From the SSAHA alignments, we identified 23,804 SNPs in 10.8 Mb of map-able sequence ($\pi = 0.0022$), of which 1,187 had the same heterozygous alleles in both the *T. b. brucei* 927 and the *T. b. gambiense* genome, indicating a false-positive rate of 5%. We identified 298 regions exhibiting higher than average diversity along the megabase chromosome. It is noteworthy, that this analysis excluded all telomere proximal regions owing to their highly repetitive nature. Whereas telomeres are well established in many species as sites of sequence variation and rearrangement [39–40], the presence of interstitial regions of high diversity in addition to the sub-telomeres is striking.

Disruptions to chromosomal colinearity are rare and reveal few subspecies-specific features

On rare occasions the otherwise consistent chromosomal colinearity is disrupted by sequence inversions and insertion-deletion events (indels). In many cases indels coincided with sequence gaps, making it difficult to confirm genuine rearrangements. Nevertheless, chromosome 10 provides two examples,

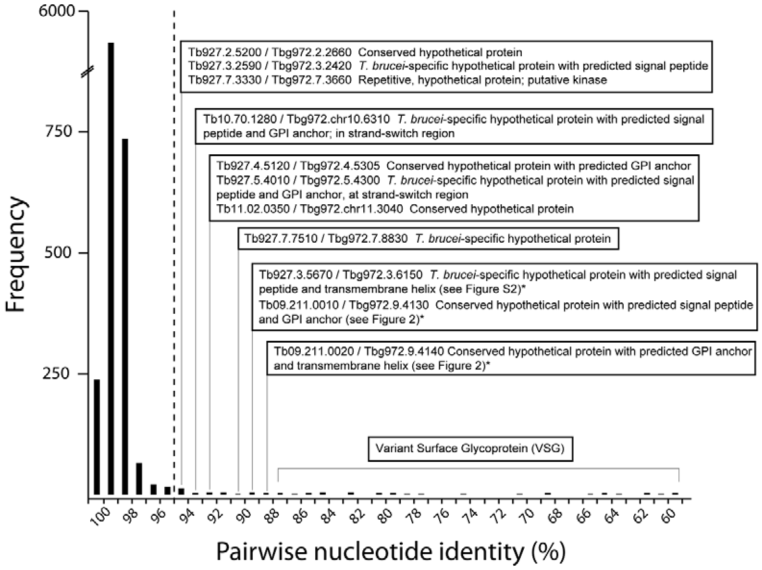


Figure 1. Frequency distribution of pairwise sequence divergence between 6929 single-copy gene orthologs in *T. b. brucei* and *T. b. gambiense*. Divergence values to the right of the dashed line are statistically significant; the identities of selected divergent gene pairs are noted. An asterisk * denotes genes belonging to a *T. b. brucei* subspecies-specific tandem gene array (see Figure 2 and Supplementary Figure S2).
doi:10.1371/journal.pntd.0000658.g001

between 275–330 kb and 3250–3350 kb, of 55 and 110 kb segmental inversions respectively. Gene order within these inverted regions remains conserved. Typically, indels have two principal causes: transposable elements and internal *VSG* 'islands'. Transposable elements such as *ingi* and RIME sequences recombine in trypanosome genomes and are responsible for several rearrangements [27]. On chromosome 9, a 7 kb insertion occurs in *T. b. brucei* due to an *ingi* element (at 1.24 Mb) not present in *T. b. gambiense*. Similarly, a 29 kb indel follows Tb11.02.5830 where an expression site-associated gene (ESAG) and a trans-sialidase gene have been inserted into *T. b. gambiense* at the corresponding position to a RIME sequence in *T. b. brucei*. By their nature, such rearrangements frequently occur in repetitive regions of the genome and, consequently, are difficult to resolve in genome assemblies. This therefore does not preclude that further events will be identified in the future.

Another source of genomic variation concerns core chromosomal *VSG* and *ESAG* genes. *VSG* genes are predominantly found in subtelomeric arrays, on intermediate or mini-chromosomes [27,13]. In addition, *VSG/ESAG* genes are less commonly found non-telomerically as 'islands', often on the opposing strand to neighbouring loci. These genes (or pseudogenes) may be: (i) atypical VSGs that do not encode all elements for accurate folding or post-translational modification; (ii) *VR* genes; or (iii) canonical *VSG* genes, imported from the subtelomere or mini-chromosome through segmental duplication. An example of the latter is a segmental insertion including 8 *VSG* genes that affects chromosome 9 in *T. b. gambiense* (Tbg972.2.570–640), since the *VSG* sequences are unrelated to each other and therefore, have not resulted from recent tandem duplications. In total, 17 such *VSG/ESAG* islands were noted in both genomes, only 6 of which were unique to one subspecies or other, including a segmental duplication in *T. b. brucei* of an atypical *VSG* combined with an insertion or deletion of *ESAG*s (Supplementary Fig. S1). Clearly, *VSG/ESAG* islands are among the more dynamic features of core chromosomes, yet where they are conserved between *T. b. brucei* and *T. b. gambiense* they contain orthologous gene sequences, indicating that they not exposed to frequent gene conversion processes like *VSG*s elsewhere.

Beyond transposable elements and *VSG* 'islands', other differences in gene order are caused by a class of small, putative coding sequences of unknown function (103 cases). These genes encode hypothetical proteins with a predicted length of 151–274 amino acids and which have no database matches to any experimentally characterized protein. Transcriptomic data (George Cross, Rockefeller University, unpublished data; Veitch et al., University of Glasgow, submitted) suggest that some of these putative genes are at least transcribed, although no product has yet been identified in proteomic assays to date (Aswini Panigrahi, SBRI, pers. comm.). Regardless of which genome encodes the putative gene, homologous sequences of high identity are found in the other genome at the corresponding positions, but without the open reading frame. Hence, they may be non-coding RNA genes or other non-coding conserved elements of undiscovered function. These features are annotated to ensure completeness, and they may yet reveal functional importance, but our view is that they are unlikely to produce proteins and will not be considered further.

An isolate-specific locus: a putative iron-ascorbate oxidoreductase in *T. b. brucei* 927

Our comparative analysis identified only a single coding sequence, a putative iron-ascorbate oxidoreductase (Tb09.211.4990), which is absent from the genomic repertoire of *T. b. gambiense*. We did not identify the *TgsGP* locus, which is known to be unique to *T. b. gambiense* [24] because it is located in the subtelomere and these regions were not fully assembled. However, sequence identical to *TgsGP* was identified among the unassembled reads. Thus, it is possible that other subspecies-specific genes exist within the subtelomeres that are not recorded here. Tb09.211.4990 is preceded upstream on chromosome 9 by a strand-switch region and downstream by both retrotransposon-like proteins and the splice-leader RNA tandem array. This region is conserved in *T. b. gambiense*, but the oxidoreductase is absent. The gene is absent from the more distantly related kinetoplastids *Leishmania major* and *T. cruzi*, as well as 9 out of 11 other *T. b. brucei* strains and a representative group 2 *T. b. gambiense* (STIB 386) that we examined with PCR primers specific to this oxidoreductase (data not shown). When compared phylogenetically with other iron-ascorbate oxidoreductases in *T. brucei*, (principally the tandem gene array at the right-hand terminus of chromosome 2, e.g. Tb927.2.6180), this protein is clearly structurally distinct (only 80% amino acid identity) and constitutes an evolutionarily old lineage. This suggests that Tb09.211.4990 is gained and lost at the population level, and that it provides additional functionality to *T. b. brucei* 927 and two other *T. b. brucei* strains in which it has been found.

Segmental duplications of putative membrane proteins contain subspecies-specific gene copies

The comparison of gene content did not identify widespread subspecies-specific loci, and found no obvious differences that could explain the distinct phenotypes of *T. brucei* subspecies. For example, ornithine decarboxylase, the target of eflornithine to which *T. b. gambiense* is uniquely sensitive, is present in single, diploid copy in both genomes and displays only a single non-synonymous substitution (N137I). We did, however, detect substantial variation within families of certain uncharacterized genes that could have important functional consequences. Such differences in co-linearity involve either the expansion of a single-copy gene in one subspecies to a tandem pair in the other, or a difference in the number of duplicates where there is a tandem array in both subspecies. Current methods of genome assembly tend to detect the first scenario (i.e., single copy vs. many) but have limitations in accurately quantifying copy number and in distinguishing between copy number and allelic variation. In fact, while the *number* of repeat units assembled can be arbitrary, the *variation* among tandem gene duplicates can be accurately assessed from genome sequence data for the two subspecies. In 20 cases, a single-copy feature (be it a single gene or chromosomal segment) in *T. b. gambiense* exists in multiple, tandem copies in *T. b. brucei*, while 8 cases of the converse were observed (Table 1). For the majority of these cases, the tandem duplicates were identical and the duplication did not result in any novel, unique sequence. But in 8 cases in *T. b. brucei*, the extra duplicates contained sequence variation that might represent subspecies-specific isoforms. In four additional cases, the would-be unique sequences were found among sequence reads of the apparently single-copy subspecies, indicating that it had been omitted from the assembly (marked by an asterisk in Table 1).

The genes involved in these *T. b. brucei*-specific segmental duplications are as yet uncharacterized, but their features suggest that they are potential sources of subspecies-specific factors and interesting opportunities for further research. They are evolutionarily novel since they are not conserved in either *T. cruzi* or *L. major*; several encode proteins with predicted cell surface roles; and some are among the fastest evolving of all *T. brucei* genes. For example, a tandem gene array of hypothetical genes encoding

Table 1. Species-specific segmental duplications in (a) *T. b. gambiense* and (b) *T. b. brucei.*

(a) ***T. b. gambiense*** **locus**	**Chr**	**Repeat unit in** ***T. b. brucei:***			**Description**	**Distance**	**Unique isoforms**
		Length	**Genes**	**Copies**			
Tbg972.3.6170[^]	3	1803	1	5	Hypothetical protein	0.0363	2
Tbg972.4.4370	4	3758	2	2	Glycosyltransferase	0.0592	1
					Hypothetical protein	0.00025	0
Tbg972.5.710	5	3379	1	3	Adenylate cyclase	0.1004	1*
Tbg972.5.4870	5	6483	1	2	Adenylate cyclase	0.00135	0
Tbg972.6.30	6	1462	1	2	Hypothetical protein	0.0005	0
Tbg972.6.140–170	6	7402	4	3	GPEET2 procyclin precursor	0.0227	1
					EP3-2 procyclin	0.0022	0
					EP3-3 procyclin	0.131	0
Tbg972.6.1030[#]	6	3865	2	5	PPIase cyclophilin-type isomerase	0.0068	0
					Hypothetical protein	0.1564	4
Tbg972.7.1230–1240	7	4162	2	2	Hypothetical protein	0.001	0
					Hypothetical protein	0.0584	1
Tbg972.7.6460	7	2147	1	2	Cell-cycle associated MOB1	0.008	1*
Tbg972.8.540	8	10346	1	2	Hypothetical protein	0.0003	0
Tbg972.8.740	8	6215	1	3	Vacuolar-type Ca2+–ATPase	0.0839	1*
Tbg972.8.1230	8	2137	1	5	gp63	0.0077	1*
Tbg972.8.1705	8	1623	1	2	Hypothetical protein	0.5671	1
Tbg972.8.7610	8	2650	1	2	Trans-sialidase	0.0222	0
Tbg972.9.1930–1950	9	6350	3	2	Ribosomal protein S7	0.0016	0
					Hypothetical protein	0.0037	0
					Nicotinamidase	0	0
Tbg972.9.4160–4130[†]	9	5729	3	5	Hypothetical protein	0.015	0
					Hypothetical protein	0.488	4
					Hypothetical protein	0.409	4
Tbg972.chr10.5040	10	3163	1	2	Elongation factor 2	0.0004	0
Tbg972.chr11.1620	11	1290	1	2	ESAG	0.481	1
Tbg972.chr11.8655	11	1325	1	2	Hypothetical protein	0.0183	0
Tbg972.chr11.12640	11	1310	1	2	Activated protein kinase c receptor	0.001	0
(b) ***T. b. brucei*** **locus**	**Chr**	**Repeat unit in** ***T. b. gambiense:***			**Description**	**Distance**	**Unique isoforms**
		Length	**Genes**	**Copies**			
Tb927.3.5050	3	1435	1	2	60S ribosomal protein L4	0.001	0
Tb927.5.2430	5	3029	1	2	Membrane transporter protein	0.00865	0
Tb09.160.0430–0450	9	6134	3	2	Protein kinase	0.0022	0
					Hypothetical protein	0.0043	0
					Hypothetical protein	0.0424	1*
Tb09.160.1520	9	2893	1	4	Hypothetical protein	0.0076	0
Tb10.70.4740	10	2272	1	2	Enolase	0	0
Tb11.27.0001	11	6755	1	2	Adenylate cyclase	0.004	0
Tb11.01.3390	11	6050	1	2	DNA topoisomerase II	0.1864	1*
Tb11.01.3610	11	3430	1	2	Membrane-bound acid phosphatase	0.0635	1*

*Sequence reads for the 'unique' isoform were found among the unassembled reads for both subspecies.

[^] see Supplementary Figure S1.

[#] see Supplementary Figure S2.

[†] see Figure 2.

doi:10.1371/journal.pntd.0000658.t001

cysteine-rich secretory proteins is shown in Supplementary Fig. S2; these are homologous to a single gene (Tbg972.3.6170) at the corresponding position on chromosome 3 in *T. b. gambiense*. From the relative strength of BLAST hits between homologs, it is clear that the first gene in the array and the singleton in *T. b. gambiense* are orthologs, while the additional copies in *T. b. brucei* (absent

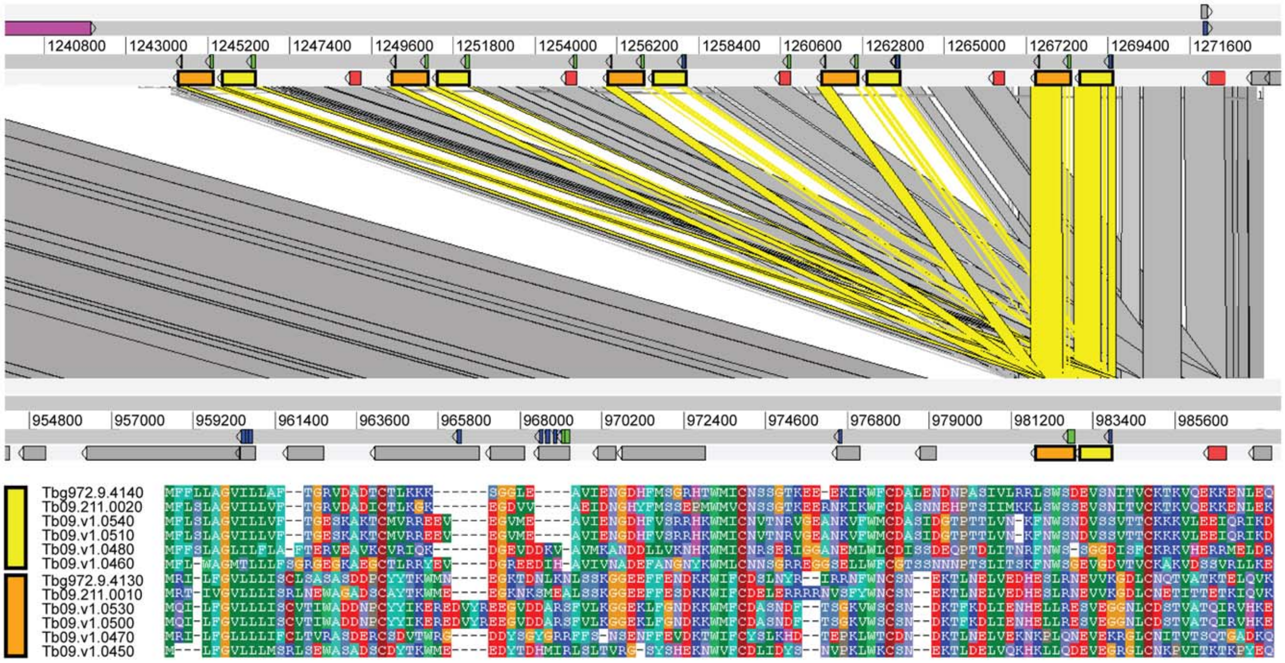


Figure 2. Segmental duplication on chromosome 9 in *T. b. brucei*. A single segment in *T. b. gambiense* comprising three coding sequences (Tbg972.9.4160, 4140 and 4130) corresponds to a three-gene segmental duplication (5 repeats) on chromosome 9 in *T. b. brucei*. The first coding sequence (shaded red) is a conserved, hypothetical gene encoding a putative secretory protein and all copies are identical. The second (shaded yellow) and third (shaded orange) coding sequences are tandem-duplicate, conserved hypothetical genes encoding putative membrane-bound proteins. Both second and third genes contain substantial sequence variation in *T. b. brucei*; the upstream-most copies are orthologous to the *T. b. gambiense* genes, but none of the remaining variants were identified among *T. b. gambiense* sequence reads. The segmental duplication is preceded immediately upstream by an INGI-mediated insertion (shaded purple).

doi:10.1371/journal.pntd.0000658.g002

from the *T. b. gambiense* read library) are unique paralogs. Indeed, they have evolved considerably, sharing only 55.1% amino acid identity with the upstream orthologs. Similarly, Figure 2 shows a single segment on chromosome 9 in *T. b. gambiense* (Tbg972.9.4160, 4140 and 4130) that corresponds to five tandem repeats in *T. b. brucei*. Among gene duplicates of the second and third coding sequences, which encode hypothetical transmembrane and GPI-anchored proteins respectively, there is considerable sequence variation (average nucleotide identities of 51.2% and 59.1% respectively). As in Supplementary Fig. S2, the 5′-most segment in *T. b. brucei* is orthologous to the *T. b. gambiense* genes, but the downstream copies are structurally divergent. A third example of segmental duplication with subsequent divergence of tandem copies occurs on chromosome 6 and concerns a hypothetical protein with a predicted signal peptide and GPI anchor (Supplementary Fig. S3).

Such segmental duplications provide rare examples of subspecies-specific gene paralogs or isoforms. It remains to be seen how common, and how ephemeral, such copy number variation is among *T. brucei* strains generally. But these cases are especially interesting because they do not simply concern gene dosage. In fact, with divergence in protein sequence often between 30–40% among paralogs, the effects on protein function could be considerable. Not only have these genes multiplied in number in very recent evolutionary time, this has been accompanied by rapid structural divergence in their predicted cell surface gene products, suggesting a role for adaptive change. Such protein isoforms could contribute to the observed differences between group 1 *T. b. gambiense* and other *T. brucei* isolates in the host-parasite relationship, both in the mammalian and insect hosts.

Tandem gene arrays frequently contain subspecies-specific sequence mosaics

Tandem gene arrays in the *T. b. brucei* genome usually contain sequence variants and analysis of tandem duplicate variation using *T. b. brucei* sequences alone showed that divergence frequently results in sequence mosaics and concerted evolution within genomes [31]. After discounting the minority of invariant tandem arrays in *T. b. gambiense*, 35 tandem gene arrays that contained sequence variation were compared with their *T. b. brucei* homologues, demonstrating that 27 arrays contained subspecies-specific gene copies (Table 2). In 5/49 instances subspecies-specific copies displayed unique sequence motifs, suggesting differential assortment of the ancestral gene repertoire between the daughter subspecies. Elsewhere, subspecies-specific copies were recombinants of other duplicates. Tests for recombination carried out on multiple alignments of gene copies from both subspecies demonstrated that sequence mosaics occurred in 31/35 data sets as exemplified by the array of invariant surface glycoproteins on chromosome 2 (ISG; Tb927.2.3270–3320) (Fig. 3). The ISG array comprises 6 and 12 gene copies in *T. b. brucei* and *T. b. gambiense*, respectively. GARD analysis detected at least five recombination breakpoints (Fig. 3a) and the recombinant nature of ISG is reflected in a highly reticulated phylogenetic network (Fig. 3b). This also identifies potential subspecies-specific recombinants, for instance, the proximity of 'Tbg7' to 'Tbg10' reflects the overall similarity of these copies, but closer inspection shows that small sections of homology exist with other copies, i.e., 'Tbg8/9'(Fig. 3c). Similarly, the intermediate position of Tbg1 reflects its affinities with multiple, unrelated sequences (Fig. 3d).

Table 2. Evidence for recombination within variable tandem gene arrays conserved in *T. b. gambiense* and *T. b. brucei.*

									Recombination tests:								
				Annotated copies:		Shared isoforms:	Unique isoforms:		PHI			LRT:			GARD:		
Array	Description	Sample identifier	Chr	*Tbb*	*Tbg*		*Tbb*	*Tbg*	Mean	Observed	p	Sequences	Sites	Breakpoints	Sequences	Sites	Breakpoints
5	Pteridine transporter	Tb927.1.2880	1	3	3	2	1	0	0.367	0.078	<0.001	6	3619	2	5	3619	9
6	Hypothetical protein	Tb927.1.4540	1	6	4	2	0	1	0.053	0.005	<0.001	9	1462	0	18	1462	8*
9	65 kDa invariant surface glycoprotein	Tb927.2.3270	2	6	10	2	0	1	0.384	0.161	<0.001	13	3226	8	10	3226	6*
12	Hypothetical protein	Tb927.2.5290	2	8	24	5	2	0	0.348	0.190	<0.001	29	1790	10	35	1790	3*
12bi	Adenosine transporter	Tb927.2.6150	2	6	3	3	3	0	0.415	0.289	<0.001	9	1407	10	9	1407	14*
12bii	Iron-ascorbate oxidoreductase	Tb927.2.6180	2	5	2	1	1	0	0.038	0.011	<0.001	4	960	0	7	960	3
13	Hypothetical protein	Tb927.3.2550	3	4	4	3	1	1	0.328	0.038	<0.001	6	1536	0	8	1536	4
17	Hypothetical protein	Tb927.3.4070	3	5	5	5	0	0	0.157	0.096	<0.001	8	2343	6	10	2343	9
22	Hypothetical protein	Tb927.4.3220	4	3	2	0	3	2	0.457	0.285	<0.001	6	2515	10	6	2515	9*
25	Adenylate cyclase, GRESAG4	Tb927.4.3880	4	3	4	3	0	1	0.179	0.062	<0.001	7	5171	2	8	5171	9*
28	Adenylate cyclase, GRESAG4	Tb927.4.4470	4	7	7	5	0	0	0.314	0.085	<0.001	13	4716	11	13	4716	4*
36	Oligosaccharyl transferase subunit	Tb927.5.890	5	3	1	1	2	0	0.391	0.274	0.001	4	2475	0	4	2475	9
41i	Procyclin	Tb927.6.450	6	3	2	2	1	0	*0.004*	*0.008*	*0.998*	*6*	*390*	*0*	*6*	*390*	*3*
42	Adenylate cyclase, GRESAG4	Tb927.6.760	6	5	4	1	1	1	0.154	0.044	<0.001	4	4701	2	9	4701	11
50	Trypanothione peroxidase 1	Tb927.7.1120	7	3	2	2	1	0	*0.232*	*0.230*	*0.467*	*5*	*510*	*1*	*5*	*510*	*5*
50bi	Hypothetical protein	Tb927.7.360	7	5	2	1	1	1	0.210	0.126	<0.001	7	1170	1	7	1170	4
55	Retrotransposon-hotspot protein	Tb927.7.2030	7	10	20	5	2	1	0.570	0.185	<0.001	25	2769	12	31	2769	3*
61	Hypothetical protein	Tb927.7.5930	7	8	6	6	0	0	0.116	0.092	<0.001	14	1755	15	14	2134	5*
62	Adenylate cyclase, GRESAG4	Tb927.7.6080	7	5	4	3	0	0	0.501	0.062	<0.001	8	5355	2	8	5355	12
62ai	Hypothetical protein	Tb927.7.6110	7	4	3	3	1	0	0.030	0.020	<0.001	7	663	1	7	663	6
62aii	Hypothetical protein	Tb927.7.6120	7	3	2	1	0	1	*0.142*	*0.166*	*0.588*	*5*	*534*	*0*	*5*	*534*	*1*
70	1,2-alpha-mannosidase IB	Tb927.8.2900	8	5	5	4	0	0	0.049	0.005	<0.001	6	1941	0	10	2020	7
75	Hypothetical protein	Tb927.8.6700	8	4	3	3	1	0	0.038	0.019	<0.001	7	2276	5	7	2276	6
80	Amino acid transporter	Tb927.8.7600	8	10	5	4	2	1	0.081	0.033	<0.001	14	1686	14	15	1686	5*
80a	Adenylate cyclase, GRESAG4	Tb927.8.7860	8	8	4	2	3	0	0.215	0.142	<0.001	10	4323	10	11	4323	11*
x2	Amino acid transporter	Tb927.8.4740	8	5	6	4	1	0	0.256	0.058	<0.001	11	1704	1	11	1704	7
84	Hypothetical protein	Tb09.160.4630	9	3	2	2	1	0	0.199	0.009	<0.001	5	1576	0	5	1576	6
93	Brucei alanine-rich protein, BARP	Tb09.244.2530	9	14	13	8	2	0	0.314	0.286	<0.001	25	1534	12	27	1534	3*
103	Hypothetical protein	Tb10.70.0040	10	4	4	3	1	1	0.078	0.061	<0.001	8	1557	4	8	1557	6
105	Hexose transporter	Tb10.6k15.2040	10	3	4	2	1	1	0.026	0.016	<0.001	7	1862	4	7	2243	7

Table 2. Cont.

								Recombination tests:									
				Annotated copies:		**Shared isoforms:**	**Unique isoforms:**		**PHI**			**LRT:**			**GARD:**		
Array	**Description**	**Sample identifier**	**Chr**	***Tbb***	***Tbg***		***Tbb***	***Tbg***	**Mean**	**Observed**	**p**	**Sequences**	**Sites**	**Breakpoints**	**Sequences**	**Sites**	**Breakpoints**
106	Procyclin-associated gene	Tb11.01.6210	10	4	2	2	2	0	0.092	0.047	<0.001	6	1185	0	6	1434	4
111	Hypothetical protein	Tb10.61.1420	10	3	3	3	0	0	0.143	0.100	<0.001	6	1102	0	6	1102	5
112	DNA polymerase kappa	Tb11.01.0080	11	10	7	4	0	0	0.332	0.167	<0.001	16	2205	9	16	2205	10*
113	Cation transporter	Tb11.01.0725	11	4	4	4	0	0	0.005	0.003	<0.001	10	1212	0	10	2458	2
126	Major surface glycoprotein, gp63	Tb11.02.5640	11	4	2	2	2	0	*0.305*	*0.396*	*1*	*4*	*1864*	*6*	*6*	*1864*	*6*

NB: Alignments of tandem gene duplicates without evidence of recombination under the PHI test are shown in italics.

*Due to data complexity, GARD analyses timed out before convergence had been achieved, returning a minimum number of breakpoint.

doi:10.1371/journal.pntd.0000658.t002

Subtelomeres are compositionally similar

Some of the hardest genome regions to reliably assemble are subtelomeres, since they usually contain numerous high-copy gene families, as well as simple and complex sequence repeats. The fluidity of subtelomeric assemblies perhaps reflects some reality about the true mutability of subtelomeric regions, since they are known to vary widely in length between trypanosome strains [41]. In comparing ~1.3 Mbp of subtelomeric sequence immediately contiguous to the chromosomal cores between the two subspecies, it is clear that they are highly similar in composition and gene order. In both *T. b. brucei* and *T. b. gambiense* the largest component of subtelomeric genes comprises *VSG*s (67.8% and 44.4%, respectively), followed by *ESAG*s (13.4%, 15.8%), and transposable element-related genes (7.6%, 13.5%). Adenylate cyclases (2.2%, 3.0%) and glycosyltransferases (1.1%, 1.5%) are also prominent features in both genomes. Beyond these subtelomeric regions, previous comparisons of telomeric VSG expression sites in various *T. brucei* strains and subspecies have established that the essential components are ubiquitous [42–43]. Hence, although *T. brucei* telomeres are known to evolve rapidly and display widespread karyotypic variation, the composition of regions beyond core chromosomes remains consistent across the species.

VSG sequence types are conserved between *T. brucei* subspecies

As the relative divergence (Fig. 1) and antigenic variability of different *T. brucei* strains is of diagnostic and clinical importance, we investigated the diversity between the *VSG* repertoires in the two subspecies by comparing all of the 1258 *VSG* sequences annotated to date in the *T. b. brucei* 927 genome with the unassembled sequence reads from the *T. b. gambiense* genome. Hence, it should be noted that we are comparing whole genes from *T. b. brucei* with fragments from *T. b. gambiense*. Among *VSG* genes with reciprocal top hits in the *T. b. gambiense* read library, the average amino acid identity is 43.3%, but with substantial variation (SD = 21.35, n = 692). Clearly, the substitution rate affecting *VSG* nucleotide sequences is relatively high, due either to positive selection, or a relaxation of purifying selection. Yet *VSG*s do not evolve so quickly as to abolish detectable orthology between subspecies; 692 *VSG* genes had a reciprocal top BLAST hit with a *T. b. gambiense* sequence read, indicating that 55% of the *T. b. brucei* repertoire (or parts thereof) were conserved in *T. b. gambiense*. Furthermore, 1061 VSG genes (84%) had reciprocal BLAST hits or very close matches, (i.e., within the top three BLAST hits for the matching *T. b. gambiense* read). The network representation emphasizes the global perspective of the VSG repertoire in *T. b. brucei* 927 relative to *T. b. gambiense* (Fig. 4). 197 VSG without close matches to *T. b. gambiense* reads are distributed throughout the network, indicating that they do not share a common origin and represent losses in *T. b. gambiense*.

As the *T. b. brucei* 927 subtelomeric sequences are incomplete, its *VSG* set is partial, and therefore further *T. b. brucei*-specific *VSG* sequences may be identified in future. That said, our analysis consistently demonstrated that *VSG* genes have corresponding sequences in both subspecies, though 787 (63%) were better related to other *T. b. brucei* genes than any *T. b. gambiense* read, suggesting that a gene duplication or gene conversion event had occurred since separation of the subspecies. We sought to identify phylogenetic structure, or discernable subsets, among VSG to establish limitations on gene conversion. The structural distinction between canonical VSG and VR proteins is already established [17] and is consistent with the location of VRs outside of the subtelomeres and their lack of pseudogenes. Accordingly, the VRs cluster together (yellow shading) in the network. They contrast

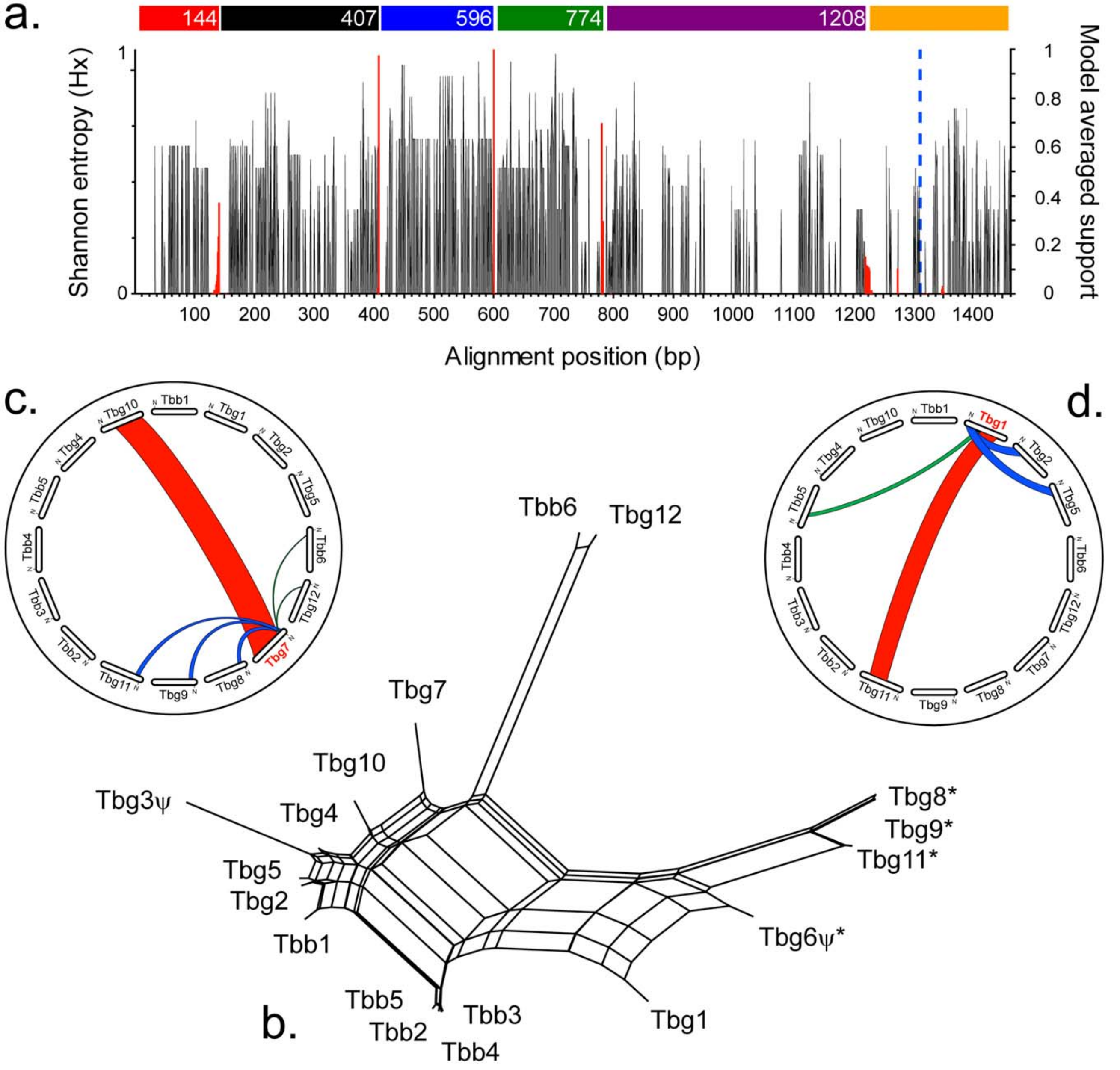


Figure 3. Analysis of sequence variation among *T. brucei* 65 kDa invariant surface glycoprotein genes. Gene copies are numbered consecutively in positional order from left to right on the chromosome, beginning with Tbb1 (Tb927.2.3270) and Tbg1 (Tbg972.2.1130) respectively. **a**. Sequence variation (expressed as Shannon entropy score, left scale) along a multiple sequence alignment, combined with recombination breakpoints (red lines, right scale) inferred by GARD analysis. A dotted blue line marks the boundary between coding and 3′ UTR regions. Coloured bars above the chart indicate recombination tracts as identified by GARD. **b**. A phylogenetic network including sequences unique to one subspecies (marked with an asterisk *). Annotated pseudogenes are indicated by ψ. **c**. A chart showing the affinities of sequence Tbg7 only; all sequences are represented in a circle, coloured bars connect Tbg7 (or regions thereof) to its closest relative among other sequences. Different colours are used to denote different affinities. **d**. Affinity chart for Tbg1.
doi:10.1371/journal.pntd.0000658.g003

with the otherwise diffuse arrangement of canonical VSG; sequences do not cluster by chromosomal location or by developmental stage expression since metacyclic-VSG are distributed throughout. VSG specific to *T. b. brucei* (red shading) do not belong to a single lineage, and equally, there is no evidence for evolutionary expansion of a particular subset. Instead, these data indicate that gene conversion has occurred frequently among subtelomeric *VSG* in the recent past, unlimited by genomic location and resulting in occasional gene loss.

The complement of variant surface glycoproteins in *T. b. gambiense* has previously been reported as being smaller, or more restricted, than that of *T. b. brucei*, based on a smaller overall genome size and reduced subtelomeric components [8]. This could indicate fewer discrete *VSG* sequence types or just fewer copies of an equal number of types. The data presented here suggest that if *T. b. gambiense* had a smaller *VSG* repertoire, this is likely to reflect quantity rather than sequence types. Hutchinson *et al.* [44] reported that, while protein sequences had diverged

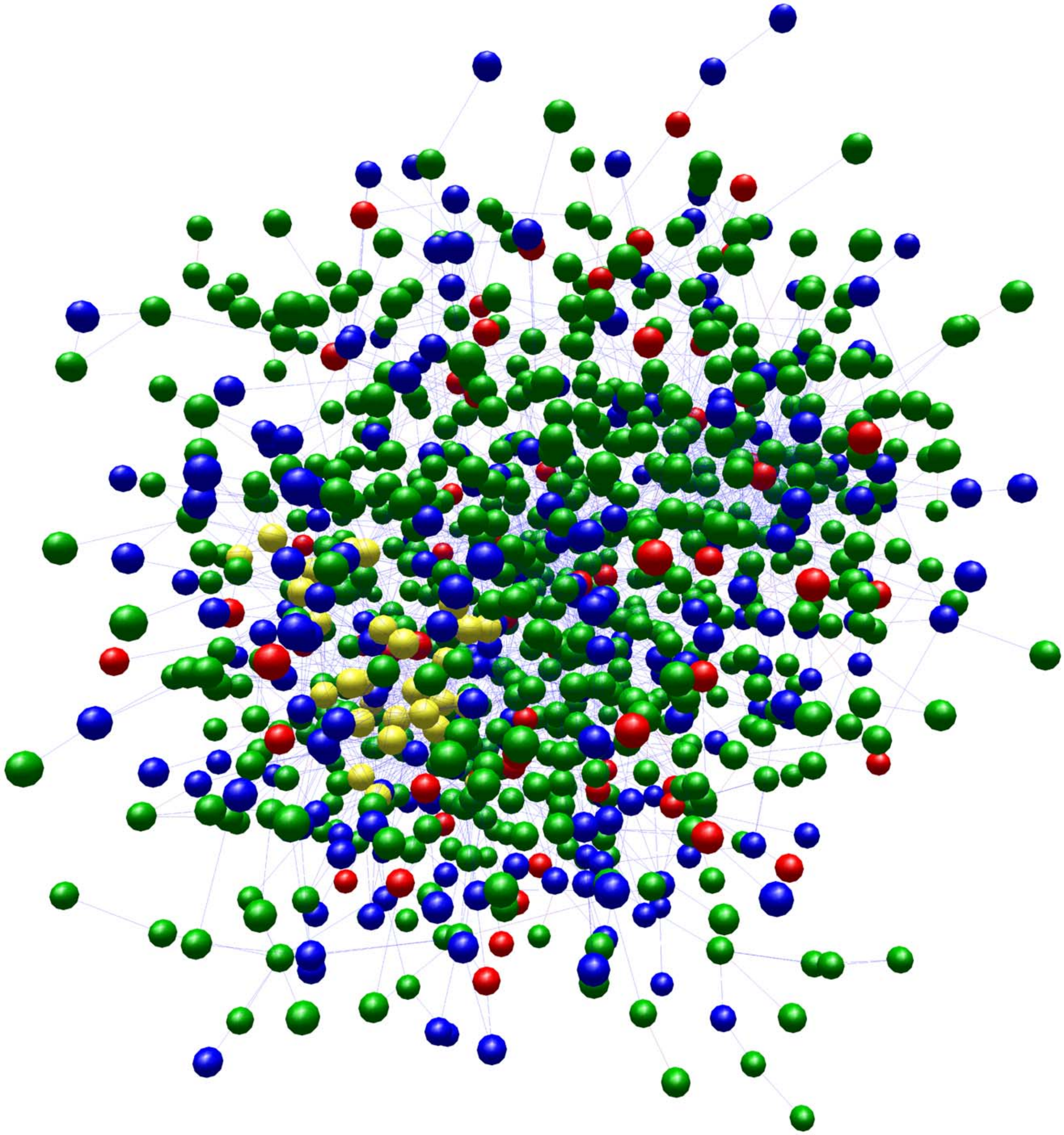


Figure 4. Variant surface glycoprotein (VSG) repertoire of *T. b. brucei* (927) represented as a three-dimensional network graph. 1258 *T. b. brucei* VSG protein sequences were compared using pairwise BLAST searches. BLAST scores were used to arrange VSG into a graph using BioLayout Express 3D 3.0. 968 individual VSG are represented as coloured spheres and are joined by edges to all other nodes with which they share >70% amino acid identity. The program minimizes the distance required to arrange all nodes such that related nodes are arranged closest to one another. Nodes are shaded by type: orthologous sequence in *T. b. gambiense* (blue), orthologous sequence in *T. b. gambiense*, but closest relative in *T. b. brucei* (green), no corresponding sequence in *T. b. gambiense* (red), metacyclic-stage VSG (purple) and VSG-related (VR) proteins (yellow).
doi:10.1371/journal.pntd.0000658.g004

consistent with subspecies-specific adaptation, 14 expressed *VSG* genes in the *T. b. brucei* Tororo strain all had close homologs in both *T. b. brucei* 927 and *T. b. gambiense*. Our data support the idea that the *VSG* repertoire is relatively stable across *T. brucei* subspecies, but that *VSG* genes have an inherently high substitution rate resulting in rapid sequence divergence relative to other genes. Corresponding *VSG* sequence types are thus likely to be found in any *T. brucei* strain, making it realistic to catalogue all *VSG* types and to monitor their expression in the field. Thus, the apparent lack of genetic hypervaribility concerning *VSG* in *T. brucei* seems simpler than other systems of antigenic variation, such as the *var* surface glycoproteins in *Plasmodium falciparum*, where frequent switching of expressed antigens is combined with genetic hypervariability and there is minimal overlap between repertoires between isolates [45–47]. Like *P. falciparum*, *T. brucei* 'shuffles its deck' with every infection, but, unlike *Plasmodium*, it always uses the same pack of cards.

Conclusion

T. b. gambiense is the most important human-infective form of *T. brucei* and currently endemic throughout central Africa. In producing a draft genome sequence for *T. b. gambiense* this study attempted to identify genetic causes for human infectivity in *T. b. gambiense*, as well as assess the scale of intraspecific genomic variation. Genomic conservation between *T. b. gambiense* and the *T. b. brucei* validates the use of *T. brucei brucei* as a model for studying the unculturable *T. b. gambiense*. Specifically, intraspecific genomic divergence is typically $<1\%$ in coding regions; gene gain and loss is associated with rare segmental duplications; indels are few and generally caused by transposable elements or *VSG/ESAG* 'islands'; and 84% of surface antigens are represented whole or in part in both subspecies. The VSG repertoire is essentially conserved at the level of modular protein domains, which are reorganized by gene conversion into novel mosaics in each strain. Therefore our data are likely to anticipate the archive present in the genomes of other strains, and a definition of total VSG diversity should be achievable through the addition of further sequences in the near future.

Comparative genomics has identified species-specific genes in other eukaryotic pathogens that display interspecific pathological variations, including *Leishmania* spp. [48]; *Candida albicans* and *C. dubliniensis* [49]; and *Plasmodium knowlesi* and *P. vivax* [50]. In applying a similar rationale here, we found no obvious candidate for a gene analogous to *SRA* that could account for human-infectivity in *T. b. gambiense*. However, since both *SRA* and *TgsGP* are homologous to *VSG* genes and subtelomeric, such a gene might not be apparent from comparison of the core chromosomes and could still exist within the subtelomeres of *T. b. gambiense*. Alternatively, rather than differences in gene content *per se*, phenotypic variability could be due to individual SNPs or indels, or to differences in gene expression. Given that innate immunity to trypanosomes in Humans is based on the uptake of high-density lipoprotein particles, which contain apoL1 and stimulate trypanolysis [20–21], perhaps the likeliest cause of phenotypic variation relates to this process. Indeed, it is possible to select for resistance to trypanolysis by down-regulating *TbHpHbR* [51], which encodes an haptoglobin-haemoglobin surface receptor that normally scavenges haem from the host, but also facilitates the uptake of trypanolytic particles [52]. However, this gene is present in *T. b. gambiense* (Tbg972.6.120) and differs from its *T. b. brucei* counterpart (Tb927.6.440) by only 5 amino acid replacements (L210S, A293V, E369G, G370A and M398I).

Hence, the basis for human infectivity in *T. b. gambiense* remains debatable, and we must now consider that features shared by both subspecies have been modified in structure or expression in *T. b. gambiense* to provide the genotypic basis for resistance to trypanolysis. This issue apart, several putative cell-surface protein families that include subspecies-specific members have been identified in *T. b. brucei*. These proteins are previously unrecognized elements of the trypanosome surface that display both recent gene duplications and accelerated evolutionary rates and we speculate that they may have acquired novel functions. We also suggest the presence or absence of such hypothetical genes varies on a population scale, and might yet contribute to phenotypic variability in host range within *T. brucei*.

Supporting Information

Figure S1 Disruption to co-linearity on chromosome 9 concerning an internal VSG 'island'.
Found at: doi:10.1371/journal.pntd.0000658.s001 (0.40 MB TIF)

Figure S2 Tandem duplication on chromosome 3 in *T. b. brucei*.
Found at: doi:10.1371/journal.pntd.0000658.s002 (0.46 MB TIF)

Figure S3 Tandem duplication on chromosome 6 in *T. b. brucei*.
Found at: doi:10.1371/journal.pntd.0000658.s003 (1.22 MB TIF)

Acknowledgments

We would like to thank the core sequencing and informatics teams at the Wellcome Trust Sanger Institute for their assistance.

Author Contributions

Conceived and designed the experiments: APJ SEM WG MB CHF. Performed the experiments: MS AB MAQ PC. Analyzed the data: APJ JM BC CHF. Contributed reagents/materials/analysis tools: MAA AM SEM WG JDB MB. Wrote the paper: APJ CHF.

References

1. Simarro PP, Jannin J, Cattand P (2008) Eliminating human African trypanosomiasis: where do we stand and what comes next? PLoS Med 5: e55.
2. World Health Organization (1998) Control and surveillance of African trypanosomiasis. Report of a WHO Expert Committee World Health Organ Tech Rep Ser.
3. Fèvre EM, Wissmann BV, Welburn SC, Lutumba P (2008) The burden of human african trypanosomiasis. PLoS Negl Trop Dis 2: e333.
4. Van Nieuwenhove S (2000) Gambiense sleeping sickness: re-emerging and soon untreatable? Bull World Health Organ 78: 1283.
5. Priotto G, Kasparian S, Mutombo W, Ngouama D, Ghorashian S, et al. (2009) Nifurtimox-eflornithine combination therapy for second-stage African *Trypanosoma brucei gambiense* trypanosomiasis: a multicentre, randomised, phase III, non-inferiority trial. Lancet 374: 56–64.
6. Lowell JE, Earl CD (2009) Leveraging biotech's drug discovery expertise for neglected diseases. Nat Biotechnol 27: 323–329.
7. Kanmogne GD, Bailey M, Gibson WC (1997) Wide variation in DNA content among isolates of *Trypanosoma brucei* ssp. Acta Trop 63: 75–87.
8. Dero B, Zampetti-Bosseler F, Pays E, Steinert M (1987) The genome and the antigen gene repertoire of *Trypanosoma brucei gambiense* are smaller than those of *T. b. brucei*. Mol Biochem Parasitol 26: 247–256.
9. Gibson WC (1986) Will the real *Trypanosoma b. gambiense* please stand up. Parasitol Today 2: 255–257.
10. Pays E, Dekerck P, Van Assel S, Babiker EA, Le Ray D, et al. (1983) Comparative analysis of a *Trypanosoma brucei gambiense* antigen gene family and its potential use in epidemiology of sleeping sickness. Mol Biochem Parasitol 7: 63–74.
11. Paindavoine P, Zampetti-Bosseler F, Pays E, Schweizer J, Guyaux M, et al. (1986) Trypanosome hybrids generated in tsetse flies by nuclear fusion. EMBO J 5: 3631–3636.
12. Gray AR (1972) Variable agglutinogenic antigens of *Trypanosoma gambiense* and their distribution among isolates of the trypanosome collected in different places in Nigeria. Trans R Soc Trop Med Hyg 66: 263–284.
13. Pays E, Salmon D, Morrison LJ, Marcello L, Barry JD (2007) Antigenic variation in *Trypanosoma brucei*, In: Barry JD, McCulloch R, Mottram J, Acosta-Serrano A, eds. Trypanosomes after the genome Horizon Bioscience. pp 339–372.
14. Borst P, Ulbert S (2001) Control of VSG gene expression sites. Mol Biochem Parasitol 114: 17–27.
15. Horn D (2004) The molecular control of antigenic variation in *Trypanosoma brucei*. Curr Mol Med 4: 563–576.
16. Taylor JE, Rudenko G (2006) Switching trypanosome coats: what's in the wardrobe? Trends Genet 22: 614–620.
17. Marcello L, Barry JD (2007) Analysis of the *VSG* gene silent archive in *Trypanosoma brucei* reveals that mosaic gene expression is prominent in antigenic variation and is favored by archive substructure. Genome Res 17: 1344–1352.
18. De Greef C, Hamers R (1994) The serum resistance-associated (*SRA*) gene of *Trypanosoma brucei* rhodesiense encodes a variant surface glycoprotein-like protein. Mol Biochem Parasitol 68: 277–284.
19. De Greef C, Chimfwembe E, Kihang'a Wabacha J, Bajyana Songa E, Hamers R (1992) Only the serum-resistant bloodstream forms of *Trypanosoma brucei rhodesiense* express the serum resistance associated (SRA) protein. Ann Soc Belg Med Trop 72 Suppl 1: 13–21.

20. Pays E, Vanhollebeke B (2008) Mutual self-defence: the trypanolytic factor story. Microbes Infect 10: 985–989.
21. Pérez-Morga D, Vanhollebeke B, Paturiaux-Hanocq F, Nolan DP, Lins L, et al. (2005) Apolipoprotein L-I promotes trypanosome lysis by forming pores in lysosomal membranes, Science 309: 469–472.
22. Xong HV, Vanhamme L, Chamekh M, Chimfwembe CE, Van Den Abbeele J, et al. (1998) A VSG expression site-associated gene confers resistance to human serum in *Trypanosoma rhodesiense*. Cell 95: 839–846.
23. Radwanska M, Chamekh M, Vanhamme L, Claes F, Magez S, et al. (2002) The serum resistance-associated gene as a diagnostic tool for the detection of *Trypanosoma brucei rhodesiense*. Am J Trop Med Hyg 67: 684–690.
24. Berberof M, Pérez-Morga D, Pays E (2001) A receptor-like flagellar pocket glycoprotein specific to *Trypanosoma brucei gambiense*. Mol Biochem Parasitol 113: 127–138.
25. Felu C, Pasture J, Pays E, Pérez-Morga D (2007) Diagnostic potential of a conserved genomic rearrangement in the *Trypanosoma brucei gambiense*-specific TGSGP locus. Am J Trop Med Hyg 76: 922–929.
26. Dukes P, Kaukus A, Hudson KM, Asonganyi T, Gashumba JK (1989) A new method for isolating *Trypanosoma brucei gambiense* from sleeping sickness patients. Trans R Soc Trop Med Hyg 83: 636–639.
27. Berriman M, Ghedin E, Hertz-Fowler C, Blandin G, Renauld H, et al. (2006) The genome of the African trypanosome *Trypanosoma brucei*. Science 309: 416–422.
28. Carver T, Berriman M, Tivey A, Patel C, Böhme U, et al. (2008) Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database. Bioinformatics 24: 2672–2676.
29. Aslett M, Aurrecoechea C, Berriman M, Brestelli J, Brunk BP, et al. (2010) TriTrypDB: a functional genomic resource for the Trypanosomatidae. Nucleic Acids Res 2010 38 (Database issue): D457–462.
30. Altshuler D, Pollara VJ, Cowles CR, Van Etten WJ, Baldwin J, et al. (2000) An SNP map of the human genome generated by reduced representation shotgun sequencing. Nature 407: 513–516.
31. Jackson AP (2007) Tandem gene arrays in *Trypanosoma brucei*: comparative phylogenomic analysis of duplicate sequence variation. BMC Evol Biol 7: 54.
32. Benz C, Nilsson D, Andersson B, Clayton C, Guilbride DL (2005) Messenger RNA processing sites in *Trypanosoma brucei*. Mol Biochem Parasitol 143: 125–134.
33. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, et al. (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23: 2947–2948.
34. Huson DH, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. Mol Biol Evol 23: 254–267.
35. Bryant D, Moulton V (2004) Neighbor-net: an agglomerative method for the construction of phylogenetic networks. Mol Biol Evol 21: 255–265.
36. Bruen TC, Philippe H, Bryant D (2006) A simple and robust statistical test for detecting the presence of recombination. Genetics 172: 2665–2681.
37. Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SD (2006) GARD: a genetic algorithm for recombination detection. Bioinformatics 22: 3096–3098.
38. Goldovsky L, Cases I, Enright AJ, Ouzounis CA (2005) BioLayout(Java): versatile network visualisation of structural and functional relationships. Appl Bioinformatics 4: 71–74.
39. Barry JD, Ginger ML, Burton P, McCulloch R (2003) Why are parasite contingency genes often associated with telomeres? Int J Parasitol 33: 29–45.
40. Scherf A, Lopez-Rubio JJ, Riviere L (2008) Antigenic variation in *Plasmodium falciparum*. Annu Rev Microbiol 62: 445–470.
41. Callejas S, Leech V, Reitter C, Melville S (2006) Hemizygous subtelomeres of an African trypanosome chromosome may account for over 75% of chromosome length. Genome Res 16: 1109–1118.
42. Hertz-Fowler C, Figueiredo LM, Quail MA, Becker M, Jackson A, et al. (2008) Telomeric expression sites are highly conserved in *Trypanosoma brucei*. PLoS One 3: e3527.
43. Young R, Taylor JE, Kurioka A, Becker M, Louis EJ, et al. (2008) Isolation and analysis of the genetic diversity of repertoires of VSG expression site containing telomeres from *Trypanosoma brucei gambiense*, *T. b. brucei* and *T. equiperdum*. BMC Genomics 9: 385.
44. Hutchinson OC, Picozzi K, Jones NG, Mott H, Sharma R, et al. (2007) Variant Surface Glycoprotein gene repertoires in *Trypanosoma brucei* have diverged to become strain-specific. BMC Genomics 8: 234.
45. Freitas-Junior LH, Bottius E, Pirrit LA, Deitsch KW, Scheidig C, et al. (2000) Frequent ectopic recombination of virulence factor genes in telomeric chromosome clusters of *P. falciparum*. Nature 407: 1018–1022.
46. Tami A, Ord R, Targett GA, Sutherland CJ (2003) Sympatric *Plasmodium falciparum* isolates from Venezuela have structured *var* gene repertoires. Malar J 2: 7.
47. Bull PC, Buckee CO, Kyes S, Kortok MM, Thathy V, et al. (2008) *Plasmodium falciparum* antigenic variation. Mapping mosaic *var* gene sequences onto a network of shared, highly polymorphic sequence blocks. Mol Microbiol 68: 1519–1534.
48. Peacock CS, Seeger K, Harris D, Murphy L, Ruiz JC, et al. (2007) Comparative genomic analysis of three *Leishmania* species that cause diverse human disease. Nat Genet 39: 839–847.
49. Jackson AP, Gamble JA, Yeomans T, Moran GP, Saunders D, et al. (2009) Comparative genomics of the fungal pathogens *Candida dubliniensis* and *C. albicans*. Genome Res Published in Advance, Sep 10, 2009.
50. Pain A, Böhme U, Berry AE, Mungall K, Finn RD, et al. (2008) The genome of the simian and human malaria parasite *Plasmodium knowlesi*. Nature 455: 799–803.
51. Faulkner SD, Oli MW, Kieft R, Cotlin L, Widener J, et al. (2006) *In vitro* generation of human high-density-lipoprotein-resistant *Trypanosoma brucei brucei*. Eukaryot Cell 5: 1276–1286.
52. Vanhollebeke B, De Muylder G, Nielsen MJ, Pays A, Tebabi P, et al. (2008) A haptoglobin-hemoglobin receptor conveys innate immunity to *Trypanosoma brucei* in humans. Science 320: 677–681.

Padmanabhan, S. and Dominiczak, A. F. (2021) Genomics of hypertension: the road to precision medicine. *Nature Reviews Cardiology*, 18(4), pp. 235-250.
(doi: 10.1038/s41569-020-00466-4)

There may be differences between this version and the published version. You are advised to consult the publisher's version if you wish to cite from it.

http://eprints.gla.ac.uk/221590/

Deposited on 1 August 2020

Enlighten – Research publications by members of the University of Glasgow
http://eprints.gla.ac.uk

Genomics of hypertension

Sandosh Padmanabhan, Anna F Dominiczak

Affiliation:
BHF Glasgow Cardiovascular Research Centre
Institute of Cardiovascular and Medical Sciences
126 University Place
University of Glasgow

Corresponding Author:
Professor Dame Anna F Dominiczak DBE MD FRCP FAHA FRSE FMedSci
Regius Professor of Medicine
Vice-Principal and Head of College of Medical, Veterinary and Life Sciences
University of Glasgow
Wolfson Medical School Building
University Avenue
Glasgow G12 8QQ

Tel No: 0141 330 2738
E-Mail: Anna.Dominiczak@glasgow.ac.uk

Abstract

The genetic architecture of blood pressure now comprises over 30 genes with rare mutations resulting in inherited forms of hypertension or hypotension and 1477 common single nucleotide polymorphisms (SNPs) associated with blood pressure. Monogenic syndromes predominantly involve the renin-angiotensin-aldosterone system, the adrenal glucocorticoid pathway and a smaller fraction due to neuroendocrine tumours of the sympathetic and parasympathetic system. The discovery of somatic mutations causing aldosterone producing adrenal adenomas revealed the critical role of calcium signalling in adrenal aldosterone synthesis. The per-SNP BP effect is small for SNPs from genome wide association studies and all the BP GWAS SNPs explain about 27% of the 30-50% estimated heritability of blood pressure. Whilst there is a paucity of GWAS SNPs mapped to known monogenic genes, a GWAS signal mapped to the uromodulin gene has been shown to affect blood pressure through sodium homeostasis, while another appears to act through endothelin. Majority of the SNPs show pleiotropic associations and unravelling these signals have the potential to understand underpinning biological pathways. In this review, we look at the current state of blood pressure genomics and explore causal pathways from mendelian randomisation studies and opportunities for drug repurposing and pharmacogenomics.

Keyword:

Hypertension; blood pressure; monogenic; repurposing; pharmacogenomics; salt; adrenal; sympathetic system

Introduction

Blood pressure is the pressure exerted by blood on walls of the arteries and is critical to the maintenance of oxygen and nutrient delivery and cardiovascular homeostasis.[1] Physiological regulatory mechanisms maintain blood pressure within a 'normal range' with lower and higher values beyond this range resulting in deleterious consequences – ischemia, myocardial infarction, stroke, kidney disease.[2] Blood pressure is normally distributed in the general population exhibiting a log-linear incremental risk of cardiovascular disease with every 1 mmHg rise in blood pressure.[3] The transition from 'normal' blood pressure to hypertension is currently based on an absolute blood pressure value (that has progressively moved downwards over the last 50 years[2,4]) at the upper end of the distribution of BP "at which the benefits of action (i.e., therapeutic intervention) exceed those of inaction".[5] Hypertension is recognised as the leading modifiable risk factor responsible for the global burden of cardiovascular disease and disability with an estimated 1.5 billion people are affected, and 10.7 million deaths (33.2% of deaths due to all risk factors) in 2015 directly attributable to a systolic BP >140 mm Hg.[6]

Though blood pressure can be parsimoniously defined as the product of cardiac output and peripheral arterial resistance, these two parameters are in turn regulated by a complex network of renal, neural, cardiac, vascular, and endocrine mechanisms under the influence of genetic and environmental factors (Figure 1). Thus, blood pressure is a multifactorial polygenic trait and by extension, hypertension, which is a dichotomisation of the quantitative blood pressure trait is likewise a multifactorial complex trait. This has been aptly illustrated in the 1960's Page model of hypertension and updated in 2014.[7,8] Systolic blood pressure shows a linear age related rise suggesting that increasing blood pressure incites vascular injury or dysfunction, which in turn leads to a vicious cycle of accumulating vascular injury and higher blood pressures.[9] The age-related rise in blood pressure is observed only in industrialised societies and not in non-westernised-tribal populations where routine dietary salt intake and other environmental stressors and exposures that increase BP are absent.[10,11] While there are clear environmental and occupational factors in addition to salt intake that influence risk of hypertension,[12] there is now accruing evidence for considerable genetic contribution from monogenic syndromes of high and low blood pressure and the exponential rate of identification of validated blood pressure single nucleotide polymorphisms (SNPs) from genome wide association studies (GWAS).[13-15] The causal genetic mutations have been established as necessary and sufficient to cause the blood pressure phenotypes in monogenic syndromes. In contrast, GWAS SNPs may provide the necessary conditions but, by themselves, do not provide sufficient cause for the manifestation of the blood pressure phenotype. In this review, we provide a broad perspective on how genomics has transformed our understanding of blood pressure and hypertension and provide a context for the opportunities and challenges for their full clinical potential to be realised. A detailed description of all the genetic variants and molecular mechanisms is not provided as these areas have been reviewed recently.[13]

Genetic contribution to blood pressure and hypertension

There is now substantial support from multiple lines of observational evidence that indicate genetics play a non-trivial role in the regulation of blood pressure. These are summarised below.

- Family and twin studies indicate a substantial heritable component for blood pressure with heritabilities ranging from 15% to 40% for clinic systolic blood pressure(SBP), and 15% to 30% for clinic diastolic blood pressure (DBP), 69% and 51% for ambulatory SBP and DBP respectively.[16,17] Heritability is a property of the population studied and whilst most of the estimates are obtained from European populations, there is data from African ancestry individuals showing similar levels of heritability.[18]
- Monozygotic twins show a greater correlation of blood pressure compared to dizygotic twins.[19] Family studies have shown risk of hypertension in an individual increases in the presence of parental and grandparental history of hypertension.[20] The familial risk of hypertension is likely to be due to genetic factors after statistical modelling showed that hypertension in grandchildren from grandparents persisted after adjustment for secular

trends, inter-generational differences in lifestyle and behaviour, physical activity and dietary sodium intake.[20]

- The existence of rare monogenic forms of high and low blood pressure with the identification of their underlying causal mutations indicates the role of specific genetic pathways primarily centred in the kidneys and adrenal glands to have a major impact on blood pressure.[13]
- The burst of discoveries in the genome wide era has unequivocally established the polygenic component in the genetic architecture of blood pressure.[14]
- Genetically determined systolic blood pressure based on 12 polymorphisms associated with blood pressure from genome wide association studies was shown to be causally associated with the age-related rise in systolic blood pressure.[21] This supports the hypothesis that increased SBP (inferred from genetic risk alleles) causes vascular injury, which in turn leads to higher SBP longitudinally.

Genetic architecture of hypertension

The genetic architecture of a trait encompasses all genetic factors and their characteristics involved in the expression of the phenotype, and is important for screening for and diagnosing disease as well as enhancing biological understanding, drug development and gene mapping.[22] It comprises the number of variants influencing a phenotype, the magnitude of their effects on the phenotype, the population frequency of these variants and their interactions with each other and the environment. Figure 2 depicts the current genetic architecture of blood pressure including 31 different genes harbouring rare causal variants for monogenic syndromes of high and low blood pressure and over 1477 common SNPs associated with blood pressure traits from GWAS.[13-15] The preponderance of common over rare variants is a reflection of the late onset nature of hypertension making it less susceptible to purifying selection. The multifactorial nature of blood pressure regulation was exemplified by the Page mosaic theory of hypertension[7] of 1960 which posited that essential hypertension (HTN) is not one disease, but several different diseases with different origins and development which included interactions among genetics, environment, adaptive, neural, mechanical, and hormonal perturbations (sympathetic nervous system, renin-angiotensin-aldosterone system). The increasing availability of large prospective population-based cohorts has enabled genetic association studies to be conducted across a wide range of phenotypes- phenome-wide association studies (PheWAS) revealing pervasive pleiotropy of a majority of the BP SNPs. The transition from the original Page's mosaic model (a wireframe model depicting the multifactorial nature of hypertension) through an update in 2014 with results from early GWAS studies to its latest iteration in Figure 2 showcase the enormous progress that have been made in unravelling the genetic architecture of blood pressure, the eclipsing of monogenic traits by polygenic signals and the scale of integrating evidence from pleiotropic associations with other phenotypes including lifestyle and environmental influences.

Monogenic syndromes

The monogenic blood pressure syndromes all present with the same BP phenotype, but are essentially separate diseases differentiated by additional clinical or laboratory characteristics and the causative genetic mutation. Molecular and clinical details of monogenic syndromes have been reviewed in detail recently and we refer to the reader to these articles.[13,23] The details of pathways and genes involved in monogenic hypertension are presented in Figure 1 and Table 1.The monogenic variants tend to have large effects on blood pressure, around 20-50 mmHg and while not contributing significantly to the public health burden of hypertension, their value lies in helping understand the blood pressure regulatory pathways that they have uncovered. An interesting theory invoking herterozygote advantage to indicate a role for rare variations in hypertension came from a study of pathogenic alleles in SLC12A1, KCNJ1, and SLC12A3 associated with a reduction blood pressure reduced risk of hypertension in the Framingham heart study, though this is yet to be validated.[24]

The mechanisms of salt-sensitive hypertension mediated through aldosterone was elucidated through the mutations causing glucocorticoid-remediable aldosteronism, Liddle syndrome, congenital adrenal hyperplasia (deficiencies of steroid 11β-hydroxylase or 17α-hydroxylase) and apparent mineralocorticoid excess which are all characterised by low plasma renin activity, hypokalaemia, and a degree of metabolic alkalosis.[25,26] About 40% of patients with adrenal aldosterone-producing adenoma (APA) have somatic gain-of-function mutations in Potassium Inwardly Rectifying Channel Subfamily J Member 5 (*KCNJ5)* gene.[23] Around 7% of APAs are due to somatic mutations in ATPase Na+/K+ Transporting Subunit Alpha 1 (*ATP1A1)*, ATPase Plasma Membrane Ca2+ Transporting 3 (*ATP2B3)*, Calcium Voltage-Gated Channel Subunit Alpha1 D (*CACNA1D)* and Catenin Beta 1 (*CTNNB1)*.[23] Mutations in these genes are less frequent in inherited cases of primary hyperaldosteronism and this raises the possibility that HTN could be due to a multiplicity of uncommon variants.[23] The mutations in *KCNJ5, CACNA1D, ATP1A1* and *ATP2B3* highlight the crucial role of calcium signalling in autonomous adrenal aldosterone production.[27] The mechanisms underlying *CTNNB1*-mediated APA formation have not been elucidated yet.[28]

Mutations in transporters in different segments of the nephron have enhanced our understanding of sodium and electrolyte regulatory mechanisms in the kidney and their impact on blood pressure. The role of the distal convoluted tubule (DCT) in blood pressure regulation was clarified through dissection of Gordon Syndrome (pseudohypoaldosteronism type II) which causes hyperkalaemic hypertension with mild metabolic acidosis. The major sodium transporter in the DCT is Na^+/Cl^+ ion cotransporter (NCC). Gordon Syndrome is caused by mutations in two serine-threonine kinase (With No Lysine Kinase 1 and 4, *WNK1* and *WNK4*) genes, which inhibit NCC expression and sodium flux. Additionally, mutations in Cullin-3 (*CUL3*) and Kelch3 (*KLHL3*) which are involved in the ubiquitination and degradation of WNK kinases also cause Gordon syndrome. Gitelman and Bartter syndromes are forms of salt losing tubulopathies characterized by hypokalaemic alkalosis with normal to low blood pressure due to dysfunctional transepithelial electrolyte transport in the thick ascending limb of loop of Henle's loop (TAL), the distal convoluted tubule (DCT disorders), or both.[29] Classical Bartter syndrome Type I and II are due to mutations in Solute Carrier Family 12 Member 1 (*SLC12A1)* and Potassium Inwardly Rectifying Channel Subfamily J Member 1 (*KCNJ1)* genes encoding respectively the $Na^+/K^+/2Cl^-$ cotransporter 2 (NKCC2) and the renal outer medullary potassium channel (ROMK) in the TAL, while Bartter Type III and V are DCT disorders due to mutations in the chloride channel ClC-Kb (*CLCNKB*) and calcium sensing receptor (CASR) respectively. Bartter syndrome with sensorineural deafness is a combined TAL and DCT disorder due to genetic defects of chloride channels ClC-Ka (*CLCNKA*) and *CLCNKB*, and the beta-subunit of barttin (*BSND*). Antenatal Bartter syndrome is a severe form manifesting in utero with foetal polyuria and hydramnios due to mutations in *SLC12A1* and *KCNJ1* genes (NKCC2 and ROMK).[29] Recently a new form of Bartter syndrome (transient antenatal Bartter syndrome) due to mutations in the melanoma-associated antigen D2 (*MAGED2*) which maps to the X chromosome has been identified which is a combined TAL/DCT disorder presenting with polyhydramnios and salt wasting.[30] Autosomal hypertension with type E brachydactyly (HTNB) is a monogenic syndrome that causes hypertension through vascular mechanisms involving phosphodiesterase 3A leading to increased neointimal proliferation and remodelling of the arteries and neurovascular structures.[31] Phaeochromocytomas and paragangliomas (PCC/PGLs) are rare neuroendocrine tumours of the adrenal glands and the sympathetic and parasympathetic paraganglia that cause hypertension through catecholamine hypersecretion. PCC/PGLs are associated with 12 genetic syndromes classified into three clusters pseudohypoxic, Wnt-signaling, and kinase-signaling clusters. Each cluster has unique clinical, molecular and imaging characteristics that can be inform precision medicine.[32,33] The pseudohypoxia group comprises of tricarboxylic acid (TCA) cycle and VHL/EPAS1 -related subgroups. The tricarboxylic acid (TCA) cycle–related subgroup is due to germline mutations in succinate dehydrogenase subunits *SDHA*, *SDHB*, *SDHC*, *SDHD, SDHAF2* (*SDHx*) and fumarate hydratase (*FH*). The VHL/EPAS1-related subgroup is caused by somatic and germline mutations in Von Hippel–Lindau tumour suppressor (*VHL*) and Endothelial PAS domain protein 1(*EPAS1*) genes resulting in multiple and recurrent PGLs. The Wnt signaling group consists of somatic mutations in Cold shock domain containing E1 (*CSDE1*) and Mastermind like transcriptional coactivator 3 (*MAML3*). The kinase signalling group consists of germline or somatic mutations in Ret proto-oncogene (*RET*), Neurofibromin 1 (*NF1*), Transmembrane protein 127 (*TMEM127*), MYC-associated factor X (*MAX*), and HRas proto-oncogene, GTPase (*HRAS*). Approximately, 60% of PCC/PGLs are caused by known genetic mutations. About 10% of PCCs and 35–40% of PGLs are metastatic. The predictors of metastatic disease include size (≥5–6 cm), extra-adrenal location of the primary tumour, noradrenergic/dopaminergic biochemical phenotype, mutations of the succinate dehydrogenase A and B (*SDHA/B*) genes, tumour multiplicity/recurrence, and age at first presentation (<20 years).[34]

Polygenic hypertension

GWAS have transformed our understanding of complex traits but their clinical impact has been limited. A major limitation has been the difficulty in linking SNP to causal gene and function primarily because a most of the GWAS variants are found in intergenic or intronic regions with only ~10% located in the coding sequence.[35,36] The key aspects of blood pressure GWAS are summarised below.

(1) All the BP GWAS SNPs explain about 27% of the 30-50% estimated heritability of blood pressure and explain about 5.7% of the phenotypic variance of SBP.[14]

(2) The effect size of any SNP on blood pressure is small, approximately 1 mmHg for SBP and 0.5 mmHg for DBP;

(3) GWAS SNPs are tag-SNPs, which means they are likely not causal SNPs, but correlate with the functional variant elsewhere.

(4) Majority of the BP loci are likely to be in novel genes and in blood pressure GWAS with a lower likelihood of identifying signals among known monogenic genes.

(5) Majority of the SNPs identified in GWAS are common with allele frequencies >1% in the population.

(6) The GWAS signals are predominantly from studies of European ancestry subjects with other ancestries under-represented and limiting the generalisability of the findings.

(7) GWAS SNP arrays do not capture rare variants and structural variations in the genome. Information is also limited from regions of low linkage disequilibrium in the genome.

(8) The utility of incorporating all variants into a genetic risk score (GRS) for individual risk prediction is still not established and there is a potential for miss-estimation of risks in non-European ancestries.[37] A one-standard deviation increase in GRS score was associated with a 3.9, 2.4 and 2.6 mmHg increase in SBP respectively among European, African and South-Asian descent individuals in the UK BioBank;[14]

(9) Large scale collaborations using whole genome sequencing in multi-ethnic populations such as Trans-Omics for Precision Medicine (https://www.nhlbi.nih.gov/science/trans-

omics-precision-medicine-topmed-program) are designed to fill the gaps in knowledge base of genomic and other omic markers of hypertension and cardiovascular disease with a view to accelerating precision medicine.

Novel pathways from GWAS

A 5'-promoter SNP, rs13333226, near the uromodulin gene (*UMOD*) was discovered from a GWAS of blood pressure extremes.[38] This SNP affects urinary uromodulin levels the product of *UMOD* which exclusively expressed in the thick ascending limb of the loop of Henle (TAL) in the kidney where 25% of the filtered sodium is reabsorbed. Uromodulin knock-out mice have low blood pressure, absent urinary uromodulin and absent salt-induced rise in BP (shift to the left of the pressure-natriuresis curve).[39] This has identified a novel pathway of blood pressure and renal function regulation through possible interaction with the Na^+-K^+-$2Cl^-$ cotransporter (NKCC2) in the TAL. NKCC2 is blocked by the commonly used loop-diuretic furosemide, which was shown by Trudu et al[40] to significantly enhance natriuresis and reduce BP levels both in the transgenic mice and in the hypertensive individuals homozygous for the UMOD increasing allele. This is now the basis of a clinical trial (www.clinicaltrials.gov NCT03354897) to reposition a loop diuretic in the hypertension care pathway. Three SNPs from BP GWAS are related to the endothelin pathway – rs9349379 (*PHACTR1*), rs1630736 (*EDN1*), rs10305838 (*EDNRA*). The minor allele of the intronic SNP, rs9349379, in the phosphatase and actin regulatory protein 1 (*PHACTR1*) gene associated with increased risk of coronary artery disease but decreased risk of migraine, cervical artery dissection, fibromuscular dysplasia, and hypertension. This SNP subsequently has been shown to be a distal regulator of *EDN1* expression.[41,42] Endothelin 1 (ET-1) is the most abundant endothelin isoform which acts in a paracrine manner on vascular smooth muscle cells (VSMC) to produce its potent vasoconstrictor effect mediated via ET subtype A (ET_A) and ET subtype B (ET_B) receptors.[43] Endothelin receptor antagonists have demonstrated a blood pressure lowering effect but did not gain traction as an antihypertensive agent because of side-effects or failure to meet clinical trial primary end-points.[43,44] There is encouraging data from phase II study with aprocitentan, a dual ET_A/ET_B endothelin receptor antagonist,[45] and this has now progressed to a phase III trial, the PRECISION study (www.clinicaltrials.gov NCT02603809). There is also evidence that endothelins play a role in non-obstructive CAD[46] and a genotype directed trial using zibotentan is currently underway (www.clinicaltrials.gov NCT04097314). Other potential targets identified include *NPR3* (natriuretic peptide receptor C) affecting blood pressure through increased vascular smooth muscle cell proliferation, angiotensin II-induced calcium flux and contraction;[47] Two other loci that are of interest include the Sodium Bicarbonate Cotransporter, Member 7 (*SLC4A7)* and Solute Carrier Family 39 (Zinc Transporter), Member 8 (*SLC39A8)* loci.[48,49] The *SLC39A8* locus is associated with an Ala391Thr variation where blood pressure increasing Ala391 variant is associated with increased ERK2 phosphorylation, NFkB activation, cadmium accumulation and reduced vascular endothelial cell viability.[49]

Polygenic risk scores

As the genetic make-up of an individual is largely stable from birth, genetic information has the potential to act as an early risk predictor. Essential HTN is influenced by multiple genetic variants with small individual effect sizes, so meaningful risk prediction necessitates examining the aggregated impact of these multiple variants through calculation of a single metric that represents an individual's overall genetic risk. Initially this was calculated as a simple genetic risk score (GRS) which is a simple additive count of the number of risk alleles (usually from a few SNPs from GWAS, sometimes weighted by effect sizes) carried by each individual.[50] More recently, with the recognition that SNPs that do not meet highly stringent significance thresholds for genome-wide association could also be predictive of disease, a broader range of SNPs, ranging from thousands to millions have been used to generate an enhanced GRS termed polygenic risk score (PRS).[51] It is important to highlight that the risk information provided by the PRS is a probabilistic range and this is different from the risk information from genetic markers of monogenic disorders, which is dichotomous (either high or low probability of disease). Additionally, the rare variant genotype points to specific biological impact of the variant while the PRS is an amalgamation of numerous small-effect variants across the genome with no specific pathway implicated. A PRS constructed using all significant BP GWAS SNPs showed a significant association with stroke, coronary artery disease, heart failure, and left ventricular mass but not with renal function.[21,52] Compared to the bottom 20% of the GRS distribution, the top 20% was associated with a 35-40% increased risk of cardiovascular disease, MI and stroke.[14] The estimated odds ratio for a 5 mmHg increase in SBP on the risk of coronary artery disease is around 1.30[1.18–1.44] supporting the hypothesis that SBP is causally related to coronary artery disease risk. This aligns with evidence from meta-analysis of clinical trials which show every 10 mm Hg reduction in systolic blood pressure significantly reduced the risk of major cardiovascular disease events by 17%-27%.[53] Lifelong genetic exposure to lower systolic blood pressure using blood pressure GWAS SNPs as genetic instruments showed participants with SBP genetic scores higher than the median had 2.9-mm Hg lower SBP and an OR of 0.82 for major coronary events (95% CI, 0.79-0.85, $P < .001$).[54] Indeed, this mendelian randomisation study confirms the independent associations of LDL-C and SBP in a dose-dependent log-linear manner with the risk of cardiovascular disease.[54] The lack of association between blood pressure PRS and renal function also suggests that progression of renal damage due to HTN may continue despite control of HTN. Whilst there is considerable interest in the use of PRS as a biomarker for early intervention or preventive strategies, this depends on its ability to motivate healthy behaviour changes for which evidence is not consistent.[55-57] A recent study from the UK BioBank showed that adherence to a healthy lifestyle (including healthy diet, limited alcohol consumption, low urinary sodium excretion, low body mass index, and increased physical activity) was associated with lower blood pressure regardless of the underlying blood pressure genetic risk suggesting that genetically predetermined rise in blood pressure and its complications can be offset at least to some extent by healthy lifestyle.[58] Prospective randomized controlled trials are essential to translate these findings into clinical practice.

Polygenic scores have mostly been developed in exclusively or majority European populations and there are concerns about the transferability of current PRS derived from European ancestry populations to other ancestries. Preliminary data indicate that PRS can still discriminate between high and low risk groups in other ethnicities,[59] but they don't perform as well, indicating clinical uses of PRS today would systematically afford greater improvement for European-descent populations and thus exacerbate health disparities.[37] The strategies to address inequities range from recalibration of scores or variant weightings for alternative populations and establishing large GWAS projects in more diverse or non-European populations.[60,61] Until these strategies yield ancestry specific data, currently the applicability of PRS will only be possible in European ancestry populations.

Pleiotropy and causality

Phenome- wide association studies (PheWAS) are similar to a genome-wide association study (GWAS), but here each genetic variant is analysed for association with a multitude of phenotypes. PheWAS permit identification of pleiotropic SNPs and overlapping traits to identify shared pathways that may help prioritise signals for follow-up studies. We performed a look-up of all the 1477 BP SNPs in GWAS catalogue and PhenoScanner to identify all non-blood pressure traits significantly associated with these SNPs using a p-value threshold of $5x10^{-5}$.[62,63] This figure encapsulates the breadth of multi-trait associations of BP SNPs from clearly plausible correlated traits to unexpected novel traits. Surprisingly, cardiovascular complications of hypertension such as coronary artery disease, stroke and atrial fibrillation show fewer pleiotropy with BP compared to adiposity, anthropometric and haematological traits.

As genotypes are determined at conception, they are independent of environmental and life-style factors and cannot be altered by disease. This means genotype variation can be leveraged as a natural experiment to establish causal relationships between risk factors and health outcomes. This is the basis for Mendelian randomisation studies.[64] The next few paragraphs look in detail at the pleiotropic traits through the lens of published mendelian randomisation studies to determine if these pleiotropic traits have a causal role in blood pressure regulation and hypertension.

Figure 3 summarises the scale of pleiotropy of all the BP SNPs from GWAS studies. Only a minority of the over 1400 SNPs are non-pleiotropic and only 6 SNPs lie in genes known to harbour mutations for monogenic blood pressure syndromes. The highest number of pleiotropic signals are for anthropometric traits which includes height, BMI, measures of adiposity and visceral fat. Next are haematological traits which include RBC, WBC and platelet measurements. Despite the high risk of stroke with hypertension, there are only 4 BP SNPs that also associate with stroke. There are 79 SNPs which also associate with CAD, and a majority of these overlap with lipid trait associations. These pleiotropic associations require further interrogation to determine if these indicate causal relationships with blood pressure or if these associations reflect confounding or reverse causation. In this context, we now summarise various mendelian randomisation studies that have tried to shed light on these associations.

Adiposity

Epidemiological studies show a correlation between BP and BMI with each 5 kg weight reduction reducing SBP by 4 mmHg.[65] In order to establish a causal role of BMI on blood pressure, one mendelian randomisation analysis used two SNPs robustly associated with BMI (rs9939609 in the Alpha-Ketoglutarate Dependent Dioxygenase (*FTO)* and rs17782313 in the Melanocortin 4 Receptor (*MC4R)* genes) and showed a 10% elevation in genotypically determined BMI was associated with a 3 mm Hg increase in systolic BP.[66] While visceral fat is known to be a better marker of cardiovascular risk, there is data to show that higher fat mass index and BMI are similarly associated with higher blood pressure suggesting that BMI may be a useful surrogate for visceral fat.[67]

Alcohol intake

Eighteen BP SNPs show pleiotropic association with alcohol intake. Interventional trials show a strong dose-response between reduction in alcohol intake and BP reduction.[68] A causal link between alcohol intake and blood pressure level has been explored using mendelian randomisation studies employing two indicator variables - rs1229984 mutation in alcohol dehydrogenase 1 (*ADH1*) and rs1229984) and rs671 in aldehyde dehydrogenase 2 (*ALDH2*) genes. The rs1229984 variant common in Caucasians heightens *ADH1* activity leading to more rapid oxidation of alcohol to acetaldehyde, while the rs671 variant common in East Asian populations inhibits metabolism of acetaldehyde by suppressing ALDH2. A large Mendelian randomisation meta-analysis of 56 epidemiological studies with 261,991 participants of European descent using the rs1229984 variant showed lower genotype-predicted alcohol intake was associated with low SBP. Similarly a study from the China Kadoorie Biobank involving 161,498 participants and using both rs1229984 and rs671 showed SBP increased by 4·8 mmHg per 280 g/week genotype-predicted alcohol intake only in males.[69,70] The gender difference may be explained by male drinkers likely consuming more per drinking session than females. While both epidemiological and mendelian randomisation results show consistent results, a caveat need to be acknowledged that ALDH2 genotype is independently associated with blood pressure as well, independent of alcohol consumption, which renders this variant unsuitable for mendelian randomisation analysis.

Birth weight

There are 28 BP SNPs associated with birth weight out of the 65 SNPs that have been associated with birth weight in genome-wide association studies (GWAS). These results suggest that there are common biological pathways that affect infant survival and future risk of cardio-metabolic diseases in adulthood.[71] The observation that higher blood pressure in later life is associated with low birth weight has two possible explanations from a mendelian randomisation analysis. It shows that maternal BP lowering alleles are associated with lower offspring birth weight (indirect maternal effect), but the same alleles inherited by the foetus result in future higher blood pressure (direct foetal effect). Birth weight-lowering genotypes in the mother does not associate with higher offspring BP implying that the inverse birth weight-blood pressure association is due to genetic effects rather than intrauterine effects.[72]

Height

Height is a pleiotropic association with 199 BP SNPs. Several epidemiological studies have reported inverse association between adult height and blood pressure (BP). One reason for high SBP in short individuals is presumed to be due to early appearance of reflected waves during systole which combines with the forward wave leading to amplification of SBP.[73] In humans, taller height is associated with reduced risk of hypertension and CAD but increased risks of atrial fibrillation, venous thromboembolism, neoplasms.[74] 194 BP SNPs also showed significant association with height. Mendelian randomisation analyses showed a one-standard deviation higher genetically determined height (~6.5 cm) results in a 16% decreased risk of CAD[75] and a 12% decreased risk of hypertension.[74] Height and lung function are closely correlated with a large number of shared genetic loci and lung function has been shown to be the mediator of the effect of height on CAD rather than blood pressure.[75]

Heart rate

Sixty-one BP SNPs show pleiotropic association with heart rate. Observational studies show that increased heart rate is a common feature in hypertensive patients and elevated heart rate is associated with the development of hypertension.[76] Mendelian randomisation studies of SNPs showed only a minor effect of resting heart rate on.[77]

Haematological traits

Blood cells are essential for oxygen transport, haemostasis, innate and acquired immune responses, iron homeostasis, vascular and endothelial function. Observational studies show and association of associations of total white blood cell, granulocyte, and neutrophil counts and RBC traits with CHD risk.[78-84] Mendelian randomisation studies indicate that there is no causal link between red blood cell count, total white blood cell, granulocyte, and neutrophil counts with CHD risk.[85] Interestingly, there is evidence for a weak positive association of CHD risk with reticulocyte indices and the weak inverse association of CHD risk with platelet volume.[85] Reticulocyte count showed a weak direct association with CHD risk,[85] and a possible explanation for this relate to higher reticulocyte count reflecting increased haemolysis with consequent higher circulating free haemoglobin which depletes NO.[86] Furthermore, free haemoglobin in blood substitutes leads to reduced nitrous oxide, increased vasoconstriction, and a higher risk of acute myocardial ischaemia.[87] Increasing platelet volume shows a weak causal relationship with lower CHD risk,[85] in contrast to systematic reviews and observational studies,[88] warranting further investigation.

Education and socio-economic status

Socioeconomic factors such as education have been associated with cardiovascular disease, though educational opportunities are not equitably distributed in the population and interventional studies are difficult.[89] A UK BioBank mendelian randomisation study showed each additional standard deviation of education (3.6 years) was associated with a 37% lower risk of coronary heart disease, and 21% of this protective effect was mediated through systolic blood pressure with BMI and smoking mediating another 21%.[90]

Alzheimer’s disease

Inherited lifetime exposure to higher SBP is associated with lower AD risk (odds ratio [OR] per standard deviation [15.4 mm Hg] of SBP [95% CI]: 0.75 [0.62–0.91]; p = 3.4 × 10−3) and with a higher probability of taking antihypertensive medication.

Pharmacogenomics

The goal of genomics is to enable precision medicine through a greater understanding of molecular pathways that regulate BP which can inform new drug development, personalisation of treatment and ultimately leading to a new taxonomy of HTN.[91] But there are significant challenges in realising this goal. Current treatment of HTN have not seen any new drug approval for over two decades. Tailoring of therapy has not progressed beyond considering self-reported African ancestry and serum renin levels.[2,4] The burgeoning list of genomic variants associated with BP and HTN opens opportunities for expanding our understanding of hypertension, revise and update the molecular taxonomy of HTN both of which will feed into hypertension precision medicine. This statement is supported by evidence from studies that showed a 6% improvement of efficacy and safety rates in drug validation efforts.[92] We connected BP GWAS signals to DrugBank[93] and Comparative Toxicogenomics Database[94] and extracted drug-gene interaction data for the plausible genes mapped to GWAS SNPs in previous studies.[14] Strikingly, all the major antihypertensive drug classes are captured by pharmacogenetic interaction with GWAS loci. This may reflect the fact that the putative published genes associated with GWAS SNPs were selected for plausible BP effect. More interestingly, gene-drug interactions revealed a large number of drugs that affect blood pressure. Many of these drugs are licensed for other conditions and have blood pressure decrease or increase as their known side-effect or adverse effect. These raise the possibility that some of the early wins from GWAS may be repurposing or repositioning opportunities for hypertension management.

Re-purposing or ‘multi-purposing’

Drug repositioning or re-purposing is the process of discovering new indications for existing drugs and is becoming increasingly relevant in the current climate of decreasing investments in new drug development and reduced rates of new drug approvals.[95] One approach that may enhance the likelihood of success is to reposition drugs against a target that has a genetic basis. Indeed, genetic studies have identified a large number of genes whose proteins are already targeted by drugs used in clinical practice (*ESR1* for tamoxifen, *CYP19A1* for aromatase inhibitors, *HMGCR* for statins) and novel drug targets (*PCSK9*).[96,97] Although GWAS SNPs still need to be causally linked to a target gene, there are significant advances in functional genomic and computational methods to accelerate discovery. Table 2 summarises other drugs with potential repurposing potential for hypertension and their pleiotropic associations. Riociguat, a guanylate cyclase (sGC) stimulator, is currently licensed for pulmonary hypertension and it is linked to the BP GWAS hit in *GUCY1A2.* Hypotension is a common side effect reported with with Riociguat, raising the possibility of potential use in blood pressure lowering.[98] Another possibility is Nesiritide is a recombinant B-type natriuretic peptide (interacts with *NPR3*, a BP GWAS locus) which failed in a clinical trial for acute decompensated heart failure and a high incidence of hypotension reported in these patients.[99] Finally, sodium-glucose cotransporter 2 (SGLT2) inhibitors are glucose lowering agents that also lower blood pressure and molecular docking studies show the SGLT2 inhibitor canagliflozin might act as a potent dual inhibitor of *ACHE* (BP GWAS locus) and SGLT2.[100]

Table 2 shows a wide range of drugs that are commonly used for other indications raising hopes for considerably increasing the pool of drugs for repurposing. However, repositioning needs to be carefully performed to ensure that patients are not exposed to off-target effects or side-effects from the wider target range of the drugs. For example, a number of anti-depressant and anti-epileptic drugs feature in Table 2 and a majority of them have hypotension as a known side-effect. Reviewing the pharmacological targets of these drugs raises caution about repositioning these drugs for hypertension. For example, olanzapine interacts with antihypertensive BP GWAS loci (*ADRB1, MTHFR-NPPA, ADRA1A, ADRA1B*), in addition its pharmacological targets include muscarinic acetylcholine receptor (CHRM1-4), histamine receptor (HRH1), 5-hydroxytryptamine receptor, kappa- and mu-type opioid receptor, Sodium-dependent serotonin transporter, HERG human cardiac K+ channel. Topiramate is an antiepileptic which interacts with the BP GWAS loci that are also linked to calcium channel antagonists (*CACNA1C, CACNA1D, CACNB2*), but in addition its pharmacological targets include gamma-aminobutyric acid receptor, voltage-gated sodium channel, carbonic anhydrase, and glutamate receptors. These two examples highlight the potential risks of repurposing drugs for hypertension, rather we suggest that the opportunities for these drugs are in 'multi-purposing' as the increasing prevalence of multimorbidity in the population requires drugs that can treat multiple conditions. Thus, a hypothetical scenario would be for patients with epilepsy and hypertension or depression and hypertension to be considered for drugs that can benefit both. Whether this strategy will improve adherence and improve outcomes for both conditions need to be assessed through randomised controlled trials.

Conclusions

Hypertension genomics is in midst of a wealth of genomic signals but a poverty of actionable results. Follow up and clinical translation of these genetic data now looks promising with integration of pleiotropic and pharmacogenomic and functional studies. The low hanging fruits, uromodulin and endothelin SNPs that have moved on to clinical trials are harbinger of accelerated advances directly from genetics to hypertension precision medicine. Polygenic risk scores look attractive, but their clinical utility needs controlled studies and the potential ethical impacts on their widespread use exacerbating health disparities need further assessment.[37,101]

Acknowledgements

We thank Dr Stefanie Lip for designing Figure 1.

Funding

Disclosures

None

References

1 Canon, W. B. Organization for physiological homeostasis. *Physiol Rev* **9**, 399-431 (1929).

2 Williams, B. *et al.* 2018 ESC/ESH Guidelines for the management of arterial hypertension. *Eur Heart J* **39**, 3021-3104, doi:10.1093/eurheartj/ehy339 (2018).

3 Lewington, S. *et al.* Age-specific relevance of usual blood pressure to vascular mortality: a meta-analysis of individual data for one million adults in 61 prospective studies. *Lancet* **360**, 1903-1913, doi:10.1016/s0140-6736(02)11911-8 (2002).

4 Whelton, P. K. *et al.* 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA Guideline for the Prevention, Detection, Evaluation, and Management of High Blood Pressure in Adults: Executive Summary: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *Circulation* **138**, e426-e483, doi:10.1161/CIR.0000000000000597 (2018).

5 Evans, J. G. & Rose, G. Hypertension. *Br Med Bull* **27**, 37-42, doi:10.1093/oxfordjournals.bmb.a070812 (1971).

6 Forouzanfar, M. H. *et al.* Global Burden of Hypertension and Systolic Blood Pressure of at Least 110 to 115 mm Hg, 1990-2015. *JAMA* **317**, 165-182, doi:10.1001/jama.2016.19043 (2017).

7 Frohlich, E. D., Dustan, H. P. & Bumpus, F. M. Irvine H. Page: 1901-1991. The celebration of a leader. *Hypertension* **18**, 443-445, doi:10.1161/01.hyp.18.4.443 (1991).

8 Padmanabhan, S., Caulfield, M. & Dominiczak, A. F. Genetic and molecular aspects of hypertension. *Circ Res* **116**, 937-959, doi:10.1161/CIRCRESAHA.116.303647 (2015).

9 Franklin, S. S. *et al.* Hemodynamic patterns of age-related changes in blood pressure. The Framingham Heart Study. *Circulation* **96**, 308-315, doi:10.1161/01.cir.96.1.308 (1997).

10 Carvalho, J. J. *et al.* Blood pressure in four remote populations in the INTERSALT Study. *Hypertension* **14**, 238-246, doi:10.1161/01.hyp.14.3.238 (1989).

11 Oliver, W. J., Cohen, E. L. & Neel, J. V. Blood pressure, sodium intake, and sodium related hormones in the Yanomamo Indians, a "no-salt" culture. *Circulation* **52**, 146-151, doi:10.1161/01.cir.52.1.146 (1975).

12 Bursztyn, M. Occupational and environmental influences on hypertension. *J Hum Hypertens* **34**, 202-206, doi:10.1038/s41371-020-0302-z (2020).

13 Padmanabhan, S. & Joe, B. Towards Precision Medicine for Hypertension: A Review of Genomic, Epigenomic, and Microbiomic Effects on Blood Pressure in Experimental Rat Models and Humans. *Physiol Rev* **97**, 1469-1528, doi:10.1152/physrev.00035.2016 (2017).

14 Evangelou, E. *et al.* Genetic analysis of over 1 million people identifies 535 new loci associated with blood pressure traits. *Nat Genet* **50**, 1412-1425, doi:10.1038/s41588-018-0205-x (2018).

15 Giri, A. *et al.* Trans-ethnic association study of blood pressure determinants in over 750,000 individuals. *Nat Genet* **51**, 51-62, doi:10.1038/s41588-018-0303-9 (2019).

16 Havlik, R. J. *et al.* Blood pressure aggregation in families. *Am J Epidemiol* **110**, 304-312, doi:10.1093/oxfordjournals.aje.a112815 (1979).

17 Kupper, N. *et al.* Heritability of daytime ambulatory blood pressure in an extended twin design. *Hypertension* **45**, 80-85, doi:10.1161/01.HYP.0000149952.84391.54 (2005).

18 Salfati, E., Morrison, A. C., Boerwinkle, E. & Chakravarti, A. Direct Estimates of the Genomic Contributions to Blood Pressure Heritability within a Population-Based Cohort (ARIC). *PLoS One* **10**, e0133031, doi:10.1371/journal.pone.0133031 (2015).

19 Luft, F. C. Twins in cardiovascular genetic research. *Hypertension* **37**, 350-356, doi:10.1161/01.hyp.37.2.350 (2001).

20 Niiranen, T. J. *et al.* Risk for hypertension crosses generations in the community: a multi-generational cohort study. *Eur Heart J* **38**, 2300-2308, doi:10.1093/eurheartj/ehx134 (2017).

21 Ference, B. A. *et al.* Clinical effect of naturally random allocation to lower systolic blood pressure beginning before the development of hypertension. *Hypertension* **63**, 1182-1188, doi:10.1161/HYPERTENSIONAHA.113.02734 (2014).

22 Timpson, N. J., Greenwood, C. M. T., Soranzo, N., Lawson, D. J. & Richards, J. B. Genetic architecture: the shape of the genetic contribution to human traits and disease. *Nat Rev Genet* **19**, 110-124, doi:10.1038/nrg.2017.101 (2018).

23 Funder, J. W. Primary Aldosteronism. *Hypertension* **74**, 458-466, doi:10.1161/HYPERTENSIONAHA.119.12935 (2019).

24 Ji, W. *et al.* Rare independent mutations in renal salt handling genes contribute to blood pressure variation. *Nat Genet* **40**, 592-599, doi:10.1038/ng.118 (2008).

25 Baranowski, E. S., Arlt, W. & Idkowiak, J. Monogenic Disorders of Adrenal Steroidogenesis. *Horm Res Paediatr* **89**, 292-310, doi:10.1159/000488034 (2018).

26 Ceccato, F. & Mantero, F. Monogenic Forms of Hypertension. *Endocrinol Metab Clin North Am* **48**, 795-810, doi:10.1016/j.ecl.2019.08.009 (2019).

27 Seidel, E., Schewe, J. & Scholl, U. I. Genetic causes of primary aldosteronism. *Exp Mol Med* **51**, 131, doi:10.1038/s12276-019-0337-9 (2019).

28 Tadjine, M., Lampron, A., Ouadi, L. & Bourdeau, I. Frequent mutations of beta-catenin gene in sporadic secreting adrenocortical adenomas. *Clin Endocrinol (Oxf)* **68**, 264-270, doi:10.1111/j.1365-2265.2007.03033.x (2008).

29 Seyberth, H. W., Weber, S. & Komhoff, M. Bartter's and Gitelman's syndrome. *Curr Opin Pediatr* **29**, 179-186, doi:10.1097/MOP.0000000000000447 (2017).

30 Laghmani, K. *et al.* Polyhydramnios, Transient Antenatal Bartter's Syndrome, and MAGED2 Mutations. *N Engl J Med* **374**, 1853-1863, doi:10.1056/NEJMoa1507629 (2016).

31 Maass, P. G. *et al.* PDE3A mutations cause autosomal dominant hypertension with brachydactyly. *Nat Genet* **47**, 647-653, doi:10.1038/ng.3302 (2015).

32 Dahia, P. L. Pheochromocytoma and paraganglioma pathogenesis: learning from genetic heterogeneity. *Nat Rev Cancer* **14**, 108-119, doi:10.1038/nrc3648 (2014).

33 Fishbein, L. *et al.* Comprehensive Molecular Characterization of Pheochromocytoma and Paraganglioma. *Cancer Cell* **31**, 181-193, doi:10.1016/j.ccell.2017.01.001 (2017).

34 Roman-Gonzalez, A. & Jimenez, C. Malignant pheochromocytoma-paraganglioma: pathogenesis, TNM staging, and current clinical trials. *Curr Opin Endocrinol Diabetes Obes* **24**, 174-183, doi:10.1097/MED.0000000000000330 (2017).

35 Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nat Genet* **48**, 245-252, doi:10.1038/ng.3506 (2016).

36 Mancuso, N. *et al.* Integrating Gene Expression with Summary Association Statistics to Identify Genes Associated with 30 Complex Traits. *Am J Hum Genet* **100**, 473-487, doi:10.1016/j.ajhg.2017.01.031 (2017).

37 Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat Genet* **51**, 584-591, doi:10.1038/s41588-019-0379-x (2019).

38 Padmanabhan, S. *et al.* Genome-wide association study of blood pressure extremes identifies variant near UMOD associated with hypertension. *PLoS Genet* **6**, e1001177, doi:10.1371/journal.pgen.1001177 (2010).

39 Graham, L. A. *et al.* Validation of uromodulin as a candidate gene for human essential hypertension. *Hypertension* **63**, 551-558, doi:10.1161/HYPERTENSIONAHA.113.01423 (2014).

40 Trudu, M. *et al.* Common noncoding UMOD gene variants induce salt-sensitive hypertension and kidney damage by increasing uromodulin expression. *Nat Med* **19**, 1655-1660, doi:10.1038/nm.3384 (2013).

41 Gupta, R. M. *et al.* A Genetic Variant Associated with Five Vascular Diseases Is a Distal Regulator of Endothelin-1 Gene Expression. *Cell* **170**, 522-533 e515, doi:10.1016/j.cell.2017.06.049 (2017).

42 Nikpay, M. *et al.* A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nat Genet* **47**, 1121-1130, doi:10.1038/ng.3396 (2015).

43 Dhaun, N. & Webb, D. J. Endothelins in cardiovascular biology and therapeutics. *Nat Rev Cardiol* **16**, 491-502, doi:10.1038/s41569-019-0176-3 (2019).

44 Barton, M. & Yanagisawa, M. Endothelin: 30 Years From Discovery to Therapy. *Hypertension* **74**, 1232-1265, doi:10.1161/HYPERTENSIONAHA.119.12105 (2019).

45 Iglarz, M. *et al.* Pharmacology of macitentan, an orally active tissue-targeting dual endothelin receptor antagonist. *J Pharmacol Exp Ther* **327**, 736-745, doi:10.1124/jpet.108.142976 (2008).

46 Ford, T. J. *et al.* Genetic dysregulation of endothelin-1 is implicated in coronary microvascular dysfunction. *Eur Heart J*, doi:10.1093/eurheartj/ehz915 (2020).

47 Ren, M. *et al.* The biological impact of blood pressure-associated genetic variants in the natriuretic peptide receptor C gene on human vascular smooth muscle. *Hum Mol Genet* **27**, 199-210, doi:10.1093/hmg/ddx375 (2018).

48 Ng, F. L. *et al.* Increased NBCn1 expression, Na+/HCO3- co-transport and intracellular pH in human vascular smooth muscle cells with a risk allele for hypertension. *Hum Mol Genet* **26**, 989-1002, doi:10.1093/hmg/ddx015 (2017).

49 Zhang, R. *et al.* A blood pressure-associated variant of the SLC39A8 gene influences cellular cadmium accumulation and toxicity. *Hum Mol Genet* **25**, 4117-4126, doi:10.1093/hmg/ddw236 (2016).

50 International Schizophrenia, C. *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748-752, doi:10.1038/nature08185 (2009).

51 Torkamani, A., Wineinger, N. E. & Topol, E. J. The personal and clinical utility of polygenic risk scores. *Nat Rev Genet* **19**, 581-590, doi:10.1038/s41576-018-0018-x (2018).

52 International Consortium for Blood Pressure Genome-Wide Association, S. *et al.* Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature* **478**, 103-109, doi:10.1038/nature10405 (2011).

53 Ettehad, D. *et al.* Blood pressure lowering for prevention of cardiovascular disease and death: a systematic review and meta-analysis. *Lancet* **387**, 957-967, doi:10.1016/S0140-6736(15)01225-8 (2016).

54 Ference, B. A. *et al.* Association of Genetic Variants Related to Combined Exposure to Lower Low-Density Lipoproteins and Lower Systolic Blood Pressure With Lifetime Risk of Cardiovascular Disease. *JAMA*, doi:10.1001/jama.2019.14120 (2019).

55 Khera, A. V. *et al.* Genetic Risk, Adherence to a Healthy Lifestyle, and Coronary Disease. *N Engl J Med* **375**, 2349-2358, doi:10.1056/NEJMoa1605086 (2016).

56 Knowles, J. W. *et al.* Impact of a Genetic Risk Score for Coronary Artery Disease on Reducing Cardiovascular Risk: A Pilot Randomized Controlled Study. *Front Cardiovasc Med* **4**, 53, doi:10.3389/fcvm.2017.00053 (2017).

57 Loos, R. J. Genetics: genome-wide risk profiles - will they change your life(style)? *Nat Rev Endocrinol* **7**, 252-254, doi:10.1038/nrendo.2011.41 (2011).

58 Pazoki, R. *et al.* Genetic Predisposition to High Blood Pressure and Lifestyle Factors: Associations With Midlife Blood Pressure Levels and Cardiovascular Events. *Circulation* **137**, 653-661, doi:10.1161/CIRCULATIONAHA.117.030898 (2018).

59 Marquez-Luna, C., Loh, P. R., South Asian Type 2 Diabetes, C., Consortium, S. T. D. & Price, A. L. Multiethnic polygenic risk scores improve risk prediction in diverse populations. *Genet Epidemiol* **41**, 811-823, doi:10.1002/gepi.22083 (2017).

60 Gurdasani, D., Barroso, I., Zeggini, E. & Sandhu, M. S. Genomics of disease risk in globally diverse populations. *Nat Rev Genet* **20**, 520-535, doi:10.1038/s41576-019-0144-0 (2019).

61 Chatterjee, N., Shi, J. & Garcia-Closas, M. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nat Rev Genet* **17**, 392-406, doi:10.1038/nrg.2016.27 (2016).

62 Kamat, M. A. *et al.* PhenoScanner V2: an expanded tool for searching human genotype-phenotype associations. *Bioinformatics*, doi:10.1093/bioinformatics/btz469 (2019).

63 MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res* **45**, D896-D901, doi:10.1093/nar/gkw1133 (2017).

64 Pingault, J. B. *et al.* Using genetic data to strengthen causal inference in observational research. *Nat Rev Genet* **19**, 566-580, doi:10.1038/s41576-018-0020-3 (2018).

65 Neter, J. E., Stam, B. E., Kok, F. J., Grobbee, D. E. & Geleijnse, J. M. Influence of weight reduction on blood pressure: a meta-analysis of randomized controlled trials. *Hypertension* **42**, 878-884, doi:10.1161/01.HYP.0000094221.86888.AE (2003).

66 Timpson, N. J. *et al.* Does greater adiposity increase blood pressure and hypertension risk?: Mendelian randomization using the FTO/MC4R genotype. *Hypertension* **54**, 84-90, doi:10.1161/HYPERTENSIONAHA.109.130005 (2009).

67 Bell, J. A. *et al.* Associations of Body Mass and Fat Indexes With Cardiometabolic Traits. *J Am Coll Cardiol* **72**, 3142-3154, doi:10.1016/j.jacc.2018.09.066 (2018).

68 Roerecke, M. *et al.* The effect of a reduction in alcohol consumption on blood pressure: a systematic review and meta-analysis. *Lancet Public Health* **2**, e108-e120, doi:10.1016/S2468-2667(17)30003-8 (2017).

69 Cho, Y. *et al.* Alcohol intake and cardiovascular risk factors: A Mendelian randomisation study. *Sci Rep* **5**, 18422, doi:10.1038/srep18422 (2015).

70 Holmes, M. V. *et al.* Association between alcohol and cardiovascular disease: Mendelian randomisation analysis based on individual participant data. *BMJ* **349**, g4164, doi:10.1136/bmj.g4164 (2014).

71 Horikoshi, M. *et al.* Genome-wide associations for birth weight and correlations with adult disease. *Nature* **538**, 248-252, doi:10.1038/nature19806 (2016).

72 Warrington, N. M. *et al.* Maternal and fetal genetic effects on birth weight and their relevance to cardio-metabolic risk factors. *Nat Genet* **51**, 804-814, doi:10.1038/s41588-019-0403-1 (2019).

73 London, G. M., Guerin, A. P., Pannier, B. M., Marchais, S. J. & Metivier, F. Body height as a determinant of carotid pulse contour in humans. *J Hypertens Suppl* **10**, S93-95 (1992).

74 Lai, F. Y. *et al.* Adult height and risk of 50 diseases: a combined epidemiological and genetic analysis. *BMC Med* **16**, 187, doi:10.1186/s12916-018-1175-7 (2018).

75 Marouli, E. *et al.* Mendelian randomisation analyses find pulmonary factors mediate the effect of height on coronary artery disease. *Commun Biol* **2**, 119, doi:10.1038/s42003-019-0361-2 (2019).

76 Palatini, P. Role of elevated heart rate in the development of cardiovascular disease in hypertension. *Hypertension* **58**, 745-750, doi:10.1161/HYPERTENSIONAHA.111.173104 (2011).

77 Larsson, S. C., Drca, N., Mason, A. M. & Burgess, S. Resting Heart Rate and Cardiovascular Disease. *Circ Genom Precis Med* **12**, e002459, doi:10.1161/CIRCGEN.119.002459 (2019).

78 Wheeler, J. G., Mussolino, M. E., Gillum, R. F. & Danesh, J. Associations between differential leucocyte count and incident coronary heart disease: 1764 incident cases from seven prospective studies of 30,374 individuals. *Eur Heart J* **25**, 1287-1292, doi:10.1016/j.ehj.2004.05.002 (2004).

79 Lassale, C. *et al.* Elements of the complete blood count associated with cardiovascular disease incidence: Findings from the EPIC-NL cohort study. *Sci Rep* **8**, 3290, doi:10.1038/s41598-018-21661-x (2018).

80 Shah, A. D., Denaxas, S., Nicholas, O., Hingorani, A. D. & Hemingway, H. Low eosinophil and low lymphocyte counts and the incidence of 12 cardiovascular diseases: a CALIBER cohort study. *Open Heart* **3**, e000477, doi:10.1136/openhrt-2016-000477 (2016).

81 Jae, S. Y. *et al.* Higher blood hematocrit predicts hypertension in men. *J Hypertens* **32**, 245-250, doi:10.1097/HJH.0000000000000029 (2014).

82 Paul, L. *et al.* Hematocrit predicts long-term mortality in a nonlinear and sex-specific manner in hypertensive adults. *Hypertension* **60**, 631-638, doi:10.1161/HYPERTENSIONAHA.112.191510 (2012).

83 Atsma, F. *et al.* Hemoglobin level is positively associated with blood pressure in a large cohort of healthy individuals. *Hypertension* **60**, 936-941, doi:10.1161/HYPERTENSIONAHA.112.193565 (2012).

84 Schaffer, A. *et al.* Impact of red blood cells count on the relationship between high density lipoproteins and the prevalence and extent of coronary artery disease: a single centre study [corrected]. *J Thromb Thrombolysis* **40**, 61-68, doi:10.1007/s11239-015-1174-x (2015).

85 Astle, W. J. *et al.* The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell* **167**, 1415-1429 e1419, doi:10.1016/j.cell.2016.10.042 (2016).

86 Gladwin, M. T., Crawford, J. H. & Patel, R. P. The biochemistry of nitric oxide, nitrite, and hemoglobin: role in blood flow regulation. *Free Radic Biol Med* **36**, 707-717, doi:10.1016/j.freeradbiomed.2003.11.032 (2004).

87 Natanson, C., Kern, S. J., Lurie, P., Banks, S. M. & Wolfe, S. M. Cell-free hemoglobin-based blood substitutes and risk of myocardial infarction and death: a meta-analysis. *JAMA* **299**, 2304-2312, doi:10.1001/jama.299.19.jrv80007 (2008).

88 Sansanayudh, N. *et al.* Mean platelet volume and coronary artery disease: a systematic review and meta-analysis. *Int J Cardiol* **175**, 433-440, doi:10.1016/j.ijcard.2014.06.028 (2014).

89 Davies, N. M., Dickson, M., Davey Smith, G., van den Berg, G. J. & Windmeijer, F. The Causal Effects of Education on Health Outcomes in the UK Biobank. *Nat Hum Behav* **2**, 117-125, doi:10.1038/s41562-017-0279-y (2018).

90 Carter, A. R. *et al.* Understanding the consequences of education inequality on cardiovascular disease: mendelian randomisation study. *BMJ* **365**, l1855, doi:10.1136/bmj.l1855 (2019).

91 Padmanabhan, S., Aman, A. & Dominiczak, A. F. Recent Findings in the Genetics of Blood Pressure: How to Apply in Practice or Is a Moonshot Required? *Curr Hypertens Rep* **20**, 54, doi:10.1007/s11906-018-0863-1 (2018).

92 Nelson, M. R. *et al.* The support of human genetic evidence for approved drug indications. *Nat Genet* **47**, 856-860, doi:10.1038/ng.3314 (2015).

93 Wishart, D. S. *et al.* DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res* **46**, D1074-D1082, doi:10.1093/nar/gkx1037 (2018).

94 Davis, A. P. *et al.* The Comparative Toxicogenomics Database: update 2019. *Nucleic Acids Res* **47**, D948-D954, doi:10.1093/nar/gky868 (2019).

95 Pushpakom, S. *et al.* Drug repurposing: progress, challenges and recommendations. *Nat Rev Drug Discov* **18**, 41-58, doi:10.1038/nrd.2018.168 (2019).

96 Jhamb, D., Magid-Slav, M., Hurle, M. R. & Agarwal, P. Pathway analysis of GWAS loci identifies novel drug targets and repurposing opportunities. *Drug Discov Today* **24**, 1232-1236, doi:10.1016/j.drudis.2019.03.024 (2019).

97 Tragante, V. *et al.* Druggability of Coronary Artery Disease Risk Loci. *Circ Genom Precis Med* **11**, e001977, doi:10.1161/CIRCGEN.117.001977 (2018).

98 Stasch, J. P., Pacher, P. & Evgenov, O. V. Soluble guanylate cyclase as an emerging therapeutic target in cardiopulmonary disease. *Circulation* **123**, 2263-2273, doi:10.1161/CIRCULATIONAHA.110.981738 (2011).

99 O'Connor, C. M. *et al.* Effect of nesiritide in patients with acute decompensated heart failure. *N Engl J Med* **365**, 32-43, doi:10.1056/NEJMoa1100171 (2011).

100 Rizvi, S. M. *et al.* Invokana (Canagliflozin) as a dual inhibitor of acetylcholinesterase and sodium glucose co-transporter 2: advancement in Alzheimer's disease- diabetes type 2 linkage via an enzoinformatics study. *CNS Neurol Disord Drug Targets* **13**, 447-451, doi:10.2174/18715273113126660160 (2014).

101 Minari, J., Brothers, K. B. & Morrison, M. Tensions in ethics and policy created by National Precision Medicine Programs. *Hum Genomics* **12**, 22, doi:10.1186/s40246-018-0151-9 (2018).

102 Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res* **19**, 1639-1645, doi:10.1101/gr.092759.109 (2009).

Table 1: Monogenic syndromes of blood pressure dysregulation with causal genes, key features and treatment.

Syndrome	Gene	Treatment
11β-hydroxylase deficiency	*CYP11B1*	Glucocorticoid therapy
17a-hydroxylase deficiency	*CYP17A1*	Glucocorticoid therapy, Potassium sparing diuretics
21-Hydroxylase deficiency	*CYP21A2*	Glucocorticoid therapy
3β-hydroxysteroid dehydrogenase	*HSD3B2*	Glucocorticoid therapy
Apparent Mineralocorticoid Excess (AME)	*HSD11B2*	Low sodium diet and spironolactone
Bartter syndrome	*CLCNKA* *CLCNKB* *KCNJ1* *MAGED2* *SLC12A1*	Potassium supplementation and use of cyclooxygenase inhibitors, angiotensin converting enzyme (ACE)-inhibitors and potassium sparing diuretics.
Familial Hyperaldosteronism (FH I)	*CYP11B1* *CYP11B2*	Dexamethasone
Familial Hyperaldosteronism (FH II)	Linkage to Chr 7p22, *KCNJ5*	Adrenalectomy is performed in case of APA, and medical therapy with aldosterone antagonists in case of BAH.
Gitelman syndrome	*SLC12A3* *CLCNKB*	Oral potassium and magnesium supplementation with adequate salt and water.
Hypertension and brachydactyly syndrome	*PDE3A*	Possible role for PDE3 inhibition
Hypertension exacerbation in pregnancy	*NR3C2*	Spironolactone contraindicated; sodium chloride treatment. Delivery of the foetus ameliorates hypertension
Liddle Syndrome	*SCNN1B* *SCNN1G*	Low sodium diet. Amiloride or triamteren.

Multiple endocrine neoplasia, type IIA neoplasia, type IIA	*RET*	Alpha adrenergic blockers for pheochromocytoma
Paragangliomas (PGL1-5)	*SDHA* *SDHAF2* *SDHB* *SDHC* *SDHD*	Surgery, adrenergic blockers (alpha blockade followed by beta-blockade)
Pseudohypoaldosteronism (PHA II- Gordon syndrome)	*CUL3* *KLHL3* *NR3C2* *WNK1* *WNK4*	Thiazide diuretics, prostaglandin inhibitors, alkalising agents, and potassium-binding resins. Na^+ and K^+ restricted diet
Sporadic aldosterone-producing adenoma (APA), or primary aldosteronism	*ATP1A1* *ATP2B3* *CACNA1D* *KCNJ5*	Surgery, aldosterone antagonists
von Hippel-Lindau syndrome	*VHL*	

Table 2: Pharmacologically active gene loci from genome wide association studies, their pleiotropic associations, key drug-gene interactions with their indications and blood pressure effect.

GWAS Locus	Pleiotropic Associations	Antihypertensive License	BP reduction as side effect	Effect on BP	Therapeutic context/indication
ACE	Non-pleiotropic	Angiotensin Converting Enzyme Inhibitors		↓↓↓↓	Hypertension; Heart Failure; DM Nephropathy
			Omapatrilat	↓↓	(Hypertension; Heart Failure – Failed because of ADR)
ACVR2A	Basophils, Eosinophils, Neutrophils, Renal.Function, Urate		Sotatercept	↓	(Pulmonary Hypertension – Phase II)
ADORA1	Non-pleiotropic		Adenosine	↓↓	Supraventricular tachycardias
			Pentoxifylline	↓↓	Peripheral vascular disease
ADRB1	Birth Weight		Bethanidine	↓↓	(Sympatholytic)
		Beta Blockers		↓↓↓↓	Hypertension; Angina; Arrhythmia; Heart Failure
			Dobutamine	↓↓	Inotropic support; cardiac stress testing
			Amiodarone	↓↓	Arrhythmia

AKR1A1	Basophil, Granulocytes, Height, Hematocrit, Haemoglobin, Neutrophils, Platelet traits, RBC traits, Reticulocytes, WBC		Tolrestat	↓	(Diabetes complications – Failed trials)
AKR1B10	Non-pleiotropic				(Diabetes complications – Failed trials)
BCL2	Adiposity, BMI, BMR, CAD, Glycemia, Hematocrit, RBC traits, Reticulocytes, T2DM, Visceral fat, Weight		Docetaxel	↓↓	Solid tumours
CACNA1C	Hematocrit, Hemoglobin, RBC traits		Cinnarizine	↓	Ménière's disease
		Spironolactone		↓↓↓↓	Hyperaldosteronism; Oedema; Heart Failure; Hypertension
			Drotaverine	↓↓	Antispasmodic
			Topiramate	↓	Epilepsy; Migraine
		Calcium Channel Blockers		↓↓↓↓	Angina; Hypertension
CACNA1D	Monogenic, Non- pleiotropic		Cinnarizine	↓	Ménière's disease
		Spironolactone		↓↓↓↓	Hyperaldosteronism; Oedema; Heart Failure; Hypertension
		Calcium Channel Blockers		↓↓↓↓	Angina; Hypertension

CACNB2	Non.Pleiotropic	Spironolactone		↓↓↓↓	Hyperaldosteronism; Oedema; Heart Failure; Hypertension
		Calcium Channel Blockers		↓↓↓↓	Angina; Hypertension
CHRM2	Heart Rate		Pizotifen	↓↓	Migraine
			Disopyramide	↓↓	Arrhythmia
			Cinnarizine	↓	Ménière's disease
			Acetylcholine	↓↓	
			Amitriptyline	↓↓	Depression; Neuropathic pain; Migraine
CSK	Cholesterol, Granulocytes, Hematocrit, Monocytes, Platelet.traits, RBC.traits, Renal.Function		Dasatinib	↓↓	Chronic myeloid leukaemia
CYP11B2	Height	Spironolactone		↓↓↓↓	Hyperaldosteronism; Oedema; Heart Failure; Hypertension
DBH	Non.Pleiotropic		Ascorbic acid	↓	Scurvy
DDAH1	Adiposity		Esomeprazole	↓	Peptic ulcer disease
EDNRA	CAD		Ambrisentan	↓↓	Pulmonary hypertension
ESR1	Adiposity, Height		Dobutamine	↓↓	Inotropic support; cardiac stress testing
FGR	BMI		Dasatinib	↓↓	Chronic myeloid leukaemia
FRK	Adiposity, Cholesterol, CRP, Height, LDL		Dasatinib	↓↓	Chronic myeloid leukaemia

GUCY1A2	CAD		Riociguat	↓↓	Pulmonary hypertension
HDAC7	Allergy, Asthma, Platelet.traits, Reticulocytes		Belinostat	↓	(T-cell lymphoma)
HDAC9	Adiposity, CAD, CVA				
HDAC9	Adiposity, CAD, CVA		Valproic acid	↓	Epilepsy; Bipolar Disorder; Migraine
HRH1	Adiposity, BMD, Neoplasm		Mirtazapine	↓↓	Depression
			Pizotifen	↓↓	5-HT, Muscarinic, H1, Alpha Adrenergic Antagonist
			Dimenhydrinate	↓	Vertigo
			Histamine	↓↓	
			Cinnarizine	↓	Ménière's disease
			Amitriptyline	↓↓	Depression; Neuropathic pain; Migraine
INSR	Adiposity, HDL, Height, Triglycerides, Urate, Visceral.fat		Insulin	↓	Diabetes Mellitus
KCNJ11	Adiposity, BMI, Glycemia, Height, T2DM		Diazoxide	↓↓	Hypoglycemia
LIMK1	Lung.function		Dabrafenib	↓↓	Melanoma
MTHFR-NPPB	CAD, RBC.traits, Visceral.fat	Carvedilol		↓	Hypertension; Angina; Heart Failure
NPR3	Adiposity, BMR, Height, Lung.function, Visceral.fat,		Nesiritide	↓↓	(Heart Failure – Failed clinical trial)

	Weight				
PDE10A	Non.Pleiotropic		Dipyridamole	↓↓	Adenosine deaminase and phosphodiesterase Inhibitor
			Papaverine	↓↓	(Antispasmodic)
PDE1A	Adiposity, BMR, Weight	Calcium Channel Blockers		↓↓↓↓	Angina; Hypertension
			Bepridil	↓↓	Angina(withdrawn)
PDE3A	CAD, Monogenic		Amrinone	↓	Heart Failure
PDE5A	Basophil, CAD, Granulocytes, Platelet.traits, WBC		Dipyridamole	↓↓	Antiplatelet
			Pentoxifylline	↓↓	Peripheral vascular disease
SCN10A	Heart.Rate, Neoplasm		Tetracaine	↓	Local anaesthetic
			Lidocaine	↓↓	Local anaesthetic; Ventricular arrhythmia
			Valproic acid	↓	Epilepsy; Bipolar Disorder; Migraine
			Brivaracetam	↓	Epilepsy
SCN2A	Adiposity		Zonisamide	↓↓	Epilepsy
			Tetracaine	↓	Local anaesthetic
			Valproic acid	↓	Epilepsy; Bipolar Disorder; Migraine
			Brivaracetam	↓	Epilepsy
VEGFA	Hematocrit, Hemoglobin, RBC.traits, Renal.Function, Urate	Carvedilol		↓↓↓↓	Hypertension; Angina; Heart Failure
YES1	Non.Pleiotropic		Dasatinib	↓↓	Chronic myeloid leukaemia

All pleiotropic associations of BP GWAS SNPs were extracted and categorised into groups of correlated traits. Some SNPs did not show any non-BP associations and were classified as non-pleiotropic. The genes linked to GWAS SNPs were determined by proximity to the SNP and cardiovascular plausibility. Only one gene per loci was included. Drug-gene interactions were obtained from Drug Bank and Comparative Toxicogenomics Database and drug indications were obtained from the British National Formulary and FDA labelled indications.

Legends to Figures.

Figure 1: Blood pressure regulation – blood pressure is the product of cardiac output and peripheral vascular resistance and these are regulated by a multitude of factors. This figure shows the spectrum of rare genomic mutations affecting blood pressure by perturbing specific physiological systems. Additionally, the role of polygenic variants are complex and integral to blood pressure regulation through pathways that are yet to be characterised. Blood pressure is a multifactorial trait and the circos background plot represents both the landscape of genetic variants and environmental and other factors that have a role in blood pressure regulation. The inner circle, represents the organ specific physiological pathways that affect blood pressure. The length of each segment is a relative representation of known genetic factors involved in that pathway. The outer circle represents the polygenic background and known environmental factors that influence multiple physiological pathways leading to the final blood pressure phenotype.

Figure 2: Genetic architecture of blood pressure and hypertension. The circos[102] plot shows depicts the monogenic and polygenic genetic variants identified by linkage, sequencing and genome wide association studies. Monogenic variants are large filled red circles and they are connected to their Clinical syndromes. The smaller blue, green and yellow filled circles are SNPs identified from genome wide association studies. The colour of the SNPs indicates whether the best association of the SNP was for systolic, diastolic or pulse pressure. FH – Familial hyperaldosteronism; AME – Apparent mineralocorticoid excess; PHA – Pseudohypoaldosteronim II; APA – Aldosterone producing adenoma; CAH – Congenital adrenal hyperplasia; HTNBRACH – Hypertension with brachydactyly; HSD3B2 - 3β-hydroxysteroid dehydrogenase deficiency; PGL1-5 – Paraganglioma; VHL – Von Hippel Lindau Syndrome; MEN – Multiple Endocrine Neoplasia II.

The top of the plot shows pleiotropic signals from PheWAS, which indicate lifestyle, environmental and early life influences on blood pressure linked to the location of the pleiotropic SNPs.

The histogram on the outer ring indicates the number of pleiotropic associations SNPs in that locus have. The taller the graph, the greater the number of pleiotropy.

Figure 3: Pleiotropic signals from phenome wide association studies. The Venn diagrams show the number of blood pressure SNPs that are significantly associated with other traits. The top panel is the global pleiotropic landscape of different phenotypic groups with the number of BP SNPs that overlap

with them shown within brackets. Here overarching phenotypic groups were condensed from a range of traits that fall under this category. More detailed representation of the constituents of the major phenotypic groups are represented in the bottom panel showing the contribution of individual traits and the degree of pleiotropy with blood pressure and the level of overlap between them.

Figure 1

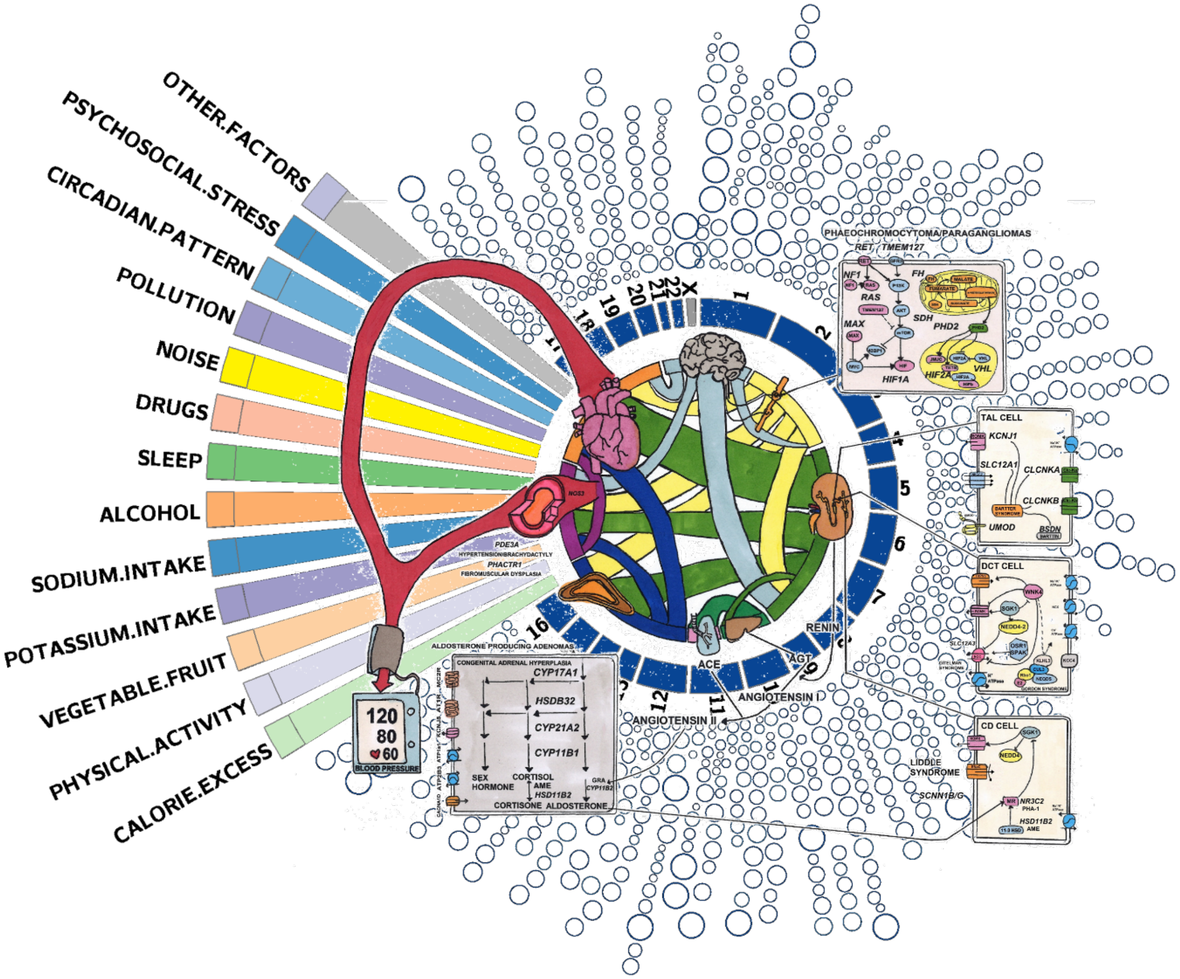
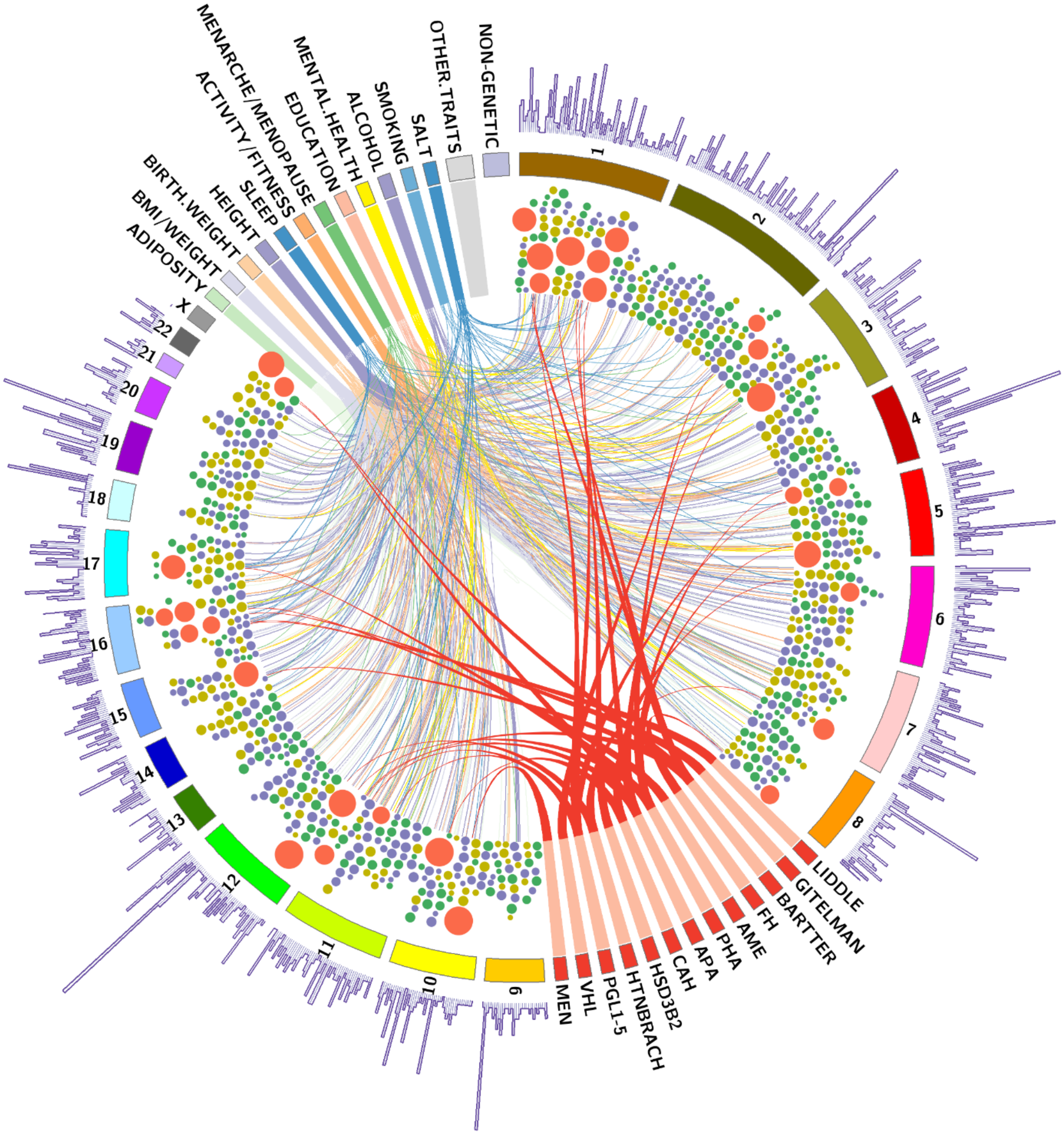
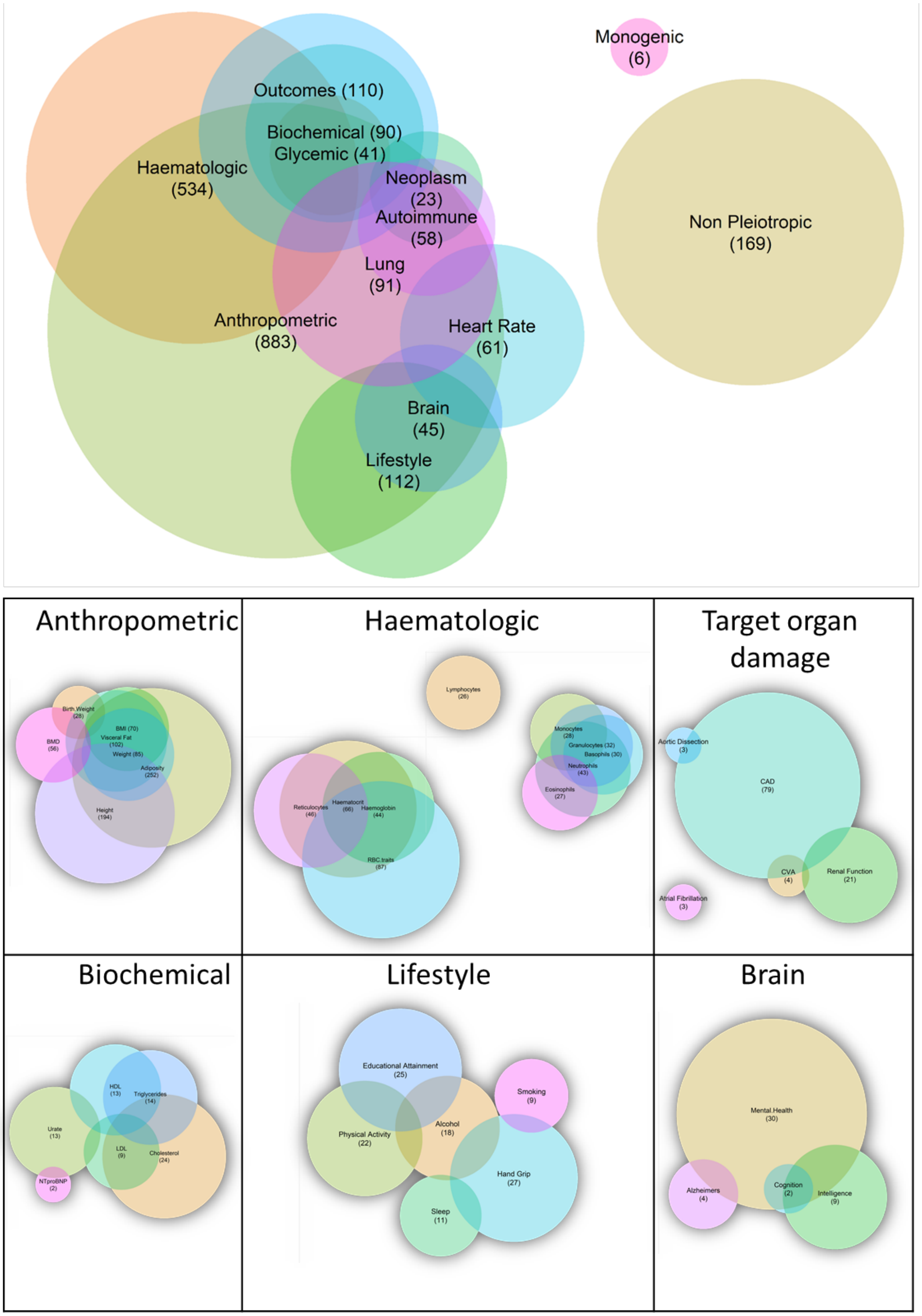

OTHER.FACTORS
PSYCHOSOCIAL.STRESS
CIRCADIAN.PATTERN
POLLUTION
NOISE
DRUGS
SLEEP
ALCOHOL
SODIUM.INTAKE
POTASSIUM.INTAKE
VEGETABLE.FRUIT
PHYSICAL.ACTIVITY
CALORIE.EXCESS
120
80
60
BLOOD PRESSURE
NOS3
PDE3A
HYPERTENSION/BRACHYDACTYLY
PHACTR1
FIBROMUSCULAR DYSPLASIA
1
2
4
5
6
7
11
12
16
17
18
19
20
21
22
X
RENIN
AGT
ACE
ANGIOTENSIN I
ANGIOTENSIN II
PHAEOCHROMOCYTOMA/PARAGANGLIOMAS
RET
TMEM127
NF1
RAS
MAX
HIF1A
FH
SDH
PHD2
VHL
HIF2A
MALATE
FUMARATE
AKT
mTOR
MYC
HIF
JMJC
TAL CELL
KCNJ1
SLC12A1
CLCNKA
CLCNKB
BARTTER SYNDROME
UMOD
BSDN
BARTTIN
DCT CELL
WNK4
SGK1
NEDD4-2
OSR1 SPAK
SLC12A3
GITELMAN SYNDROME
KLHL3
CUL3
KCC4
GORDON SYNDROME
CD CELL
SGK1
NEDD4
LIDDLE SYNDROME
SCNN1B/G
MR
NR3C2
PHA-1
HSD11B2
AME
11-β HSD
ALDOSTERONE PRODUCING ADENOMAS
CONGENITAL ADRENAL HYPERPLASIA
CYP17A1
HSDB32
CYP21A2
CYP11B1
SEX HORMONE
CORTISOL
AME
HSD11B2
GRA
CYP11B2
CORTISONE
ALDOSTERONE
CACNA1D
ATP2B3
ATP1A1
KCNJ5
AT1R
MC2R

Figure 2



NON-GENETIC
OTHER.TRAITS
SALT
SMOKING
ALCOHOL
MENTAL.HEALTH
EDUCATION
MENARCHE/MENOPAUSE
ACTIVITY/FITNESS
SLEEP
HEIGHT
BIRTH.WEIGHT
BMI/WEIGHT
ADIPOSITY
X
22
21
20
19
18
17
16
15
14
13
12
11
10
9
1
2
3
4
5
6
7
8
MEN
VHL
PGL1-5
HTNBRACH
HSD3B2
CAH
APA
PHA
AME
FH
BARTTER
GITELMAN
LIDDLE

Figure 3


Outcomes (110)
Biochemical (90)
Glycemic (41)
Haematologic (534)
Neoplasm (23)
Autoimmune (58)
Lung (91)
Anthropometric (883)
Heart Rate (61)
Brain (45)
Lifestyle (112)
Monogenic (6)
Non Pleiotropic (169)
Anthropometric
Birth Weight (28)
BMD (56)
BMI (70)
Visceral Fat (102)
Weight (85)
Adiposity (252)
Height (194)
Haematologic
Lymphocytes (26)
Monocytes (28)
Granulocytes (32)
Basophils (30)
Neutrophils (43)
Eosinophils (27)
Reticulocytes (46)
Haematocrit (66)
Haemoglobin (44)
RBC traits (87)
Target organ damage
Aortic Dissection (3)
CAD (79)
CVA (4)
Renal Function (21)
Atrial Fibrillation (3)
Biochemical
HDL (13)
Triglycerides (14)
Urate (13)
LDL (9)
Cholesterol (24)
NTproBNP (2)
Lifestyle
Educational Attainment (25)
Smoking (9)
Physical Activity (22)
Alcohol (18)
Hand Grip (27)
Sleep (11)
Brain
Mental Health (30)
Alzheimers (4)
Cognition (2)
Intelligence (9)

bioRxiv preprint first posted online Aug. 7, 2018; doi: http://dx.doi.org/10.1101/386078. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

Joint sequencing of human and pathogen genomes reveals the genetics of pneumococcal meningitis

Authors: John A. Lees[1,2], Bart Ferwerda[3], Philip H. C. Kremer[3], Nicole E. Wheeler[2], Mercedes Valls Serón[3], Nicholas J. Croucher[4], Rebecca A. Gladstone[2], Hester J. Bootsma[5], Nynke Rots[5], Alienke J. Wijmega-Monsuur[5], Elisabeth A. M. Sanders[5,6], Krzysztof Trzciński[5,6], Anne L. Wyllie[6], Aeilko H. Zwinderman[7], Leonard H. van den Berg[8], Wouter van Rheenen[8], Jan H. Veldink[8], Zitta B. Harboe[9], Lene F. Lundbo[10], Lisette C. P. G. M. de Groot[11], Natasja M. van Schoor[12], Nathalie van der Velde[13,14], Lars H. Ängquist[15], Thorkild I. A. Sørensen[16,17], Ellen A. Nohr[18], Alexander J. Mentzer[19], Tara C. Mills[19], Julian C. Knight[19], Mignon du Plessis[20], Susan Nzenze[20], Jeffrey N. Weiser[1], Julian Parkhill[2], Shabir Madhi[21], Thomas Benfield[10], Anne von Gottberg[21], Arie van der Ende[22,23], Matthijs C. Brouwer[3], Jeffrey C. Barrett[2], Stephen D. Bentley[2,†,*], Diederik van de Beek[3,†,*]

[†]Contributed equally

*Corresponding authors: Diederik van de Beek (d.vandebeek@amc.uva.nl) and Stephen Bentley (sdb@sanger.ac.uk)

Affiliations:

[1]Department of Microbiology, New York University School of Medicine, New York, United States of America

[2]Infection Genomics, Wellcome Sanger Institute, Cambridge, United Kingdom

[3]Department of Neurology, Amsterdam Academic Medical Centre, University of Amsterdam, Amsterdam, Amsterdam Neuroscience, The Netherlands

[4]Department of Infectious Disease Epidemiology, Imperial College London, London, United Kingdom

[5]Centre for Infectious Disease Control, National Institute for Public Health and the Environment, Bilthoven, The Netherlands

[6]Department of Pediatric Immunology and Infectious Diseases, Wilhemina's Children Hospital, University Medical Centre Utrecht, Utrecht, The Netherlands

[7]Department of Biostatistics, Academic Medical Center, Amsterdam, The Netherlands

[8]Department of Neurology, Brain Center Rudolf Magnus, University Medical Center Utrecht, Utrecht, The Netherlands

[9]Department of Microbiological Surveillance and Research, Statens Serum Institut, Copenhagen, Denmark

[10]Department of Infectious Diseases, Hvidovre Hospital, University of Copenhagen, Hvidovre, Denmark

[11]Department of Human Nutrition, Wageningen University, PO Box 17 6700 AA Wageningen, The Netherlands

[12]Amsterdam Public Health Research Institute, Department of Epidemiology and Biostatistics, VU University Medical Center, Van der Boechorststraat 7, 1081 BT Amsterdam, the Netherlands

bioRxiv preprint first posted online Aug. 7, 2018; doi: http://dx.doi.org/10.1101/386078. The copyright holder for this preprint
(which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license.

[13]Amsterdam Public Health Research Institute, Department of Internal Medicine, Geriatrics, AMC, Meibergdreef 9, 1105 AZ Amsterdam, The Netherlands
[14]Department of Internal Medicine, Erasmus MC, University Medical Centre Rotterdam, P.O. Box 2040, 3000 CA Rotterdam, The Netherlands
[15]Center for Clinical Research and Disease Prevention, Bispebjerg and Frederiksberg Hospitals, The Capital Region, Copenhagen, Denmark
[16]The Novo Nordisk Foundation Center for Basic Metabolic Research, Section of Metabolic Genetics, Denmark
[17]The Department of Public Health, Section of Epidemiology, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark
[18]Institute of Clinical Research, University of Southern Denmark, Odense, Denmark
[19]Wellcome Centre for Human Genetics, Nuffield Department of Medicine, University of Oxford, Oxford, United Kingdom
[20]University of Witwatersrand, Johannesburg, South Africa
[21]National Institute for Communicable Diseases, Johannesburg, South Africa
[22]Department of Medical Microbiology, Centre for Infection and Immunity Amsterdam (CINIMA), Academic Medical Centre, University of Amsterdam,
Amsterdam, The Netherlands
[23]Netherlands Reference Laboratory for Bacterial Meningitis, Academic Medical Centre/RIVM, University of Amsterdam, Amsterdam, The Netherlands

bioRxiv preprint first posted online Aug. 7, 2018; doi: http://dx.doi.org/10.1101/386078. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

Abstract

Streptococcus pneumoniae is a common nasopharyngeal colonizer, but can also cause life-threatening invasive diseases such as empyema, bacteremia and meningitis. Genetic variation of host and pathogen is known to play a role in invasive pneumococcal disease, though to what extent is unknown. In a genome-wide association study of human and pathogen we show that human variation explains almost half of variation in susceptibility to pneumococcal meningitis and one-third of variation in severity, and identified variants in *CCDC33* associated with susceptibility. Pneumococcal variation explained a large amount of invasive potential, but serotype explained only half of this variation. Newly developed methods identified pneumococcal genes involved in invasiveness including *pspC* and *zmpD*, and allowed a human-bacteria interaction analysis, finding associations between pneumococcal lineage and *STK32C*.

bioRxiv preprint first posted online Aug. 7, 2018; doi: http://dx.doi.org/10.1101/386078. The copyright holder for this preprint
(which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license.

Streptococcus pneumoniae, or the pneumococcus, is a leading cause of pneumonia, meningitis, and bacteremia. Over 90 serotypes are known, which have varying prevalence of asymptomatic carriage and disease[1,2]. Some clonal genotypes have been associated with invasive disease, though as serotype is correlated with genetic background, finding how much each of these factors affects invasive propensity in clinical cases of disease is challenging.

Bacterial meningitis involves severe inflammation of the membranes surrounding the brain, the meninges, which is a response to the presence of bacteria in the cerebrospinal fluid (CSF)[3]. *S. pneumoniae* is the most common cause of bacterial meningitis and despite advances in vaccination and treatment case fatality rate is 17-20% and unfavourable outcome occurs in 38-50% of cases[4].

Knowledge of the contribution of genetic variability of humans and invading pathogens to pneumococcal meningitis susceptibility could guide development of new vaccines preventing the progression from asymptomatic carriage to invasive disease, whereas genetic variation associated with disease severity may guide new clinical intervention strategies during treatment[5]. However the effect of human genetics on pneumococcal meningitis is unknown – whether it affects the disease at all, and if so, which specific regions of the genome cause the effect. Historically, genetic association studies on bacterial meningitis have been held back by only assessing candidate genes, small sample sizes or poorly defined phenotypes[6]. More recent GWAS studies have found associations for children with meningococcal meningitis in Europe[7], and pneumococcal meningitis in Kenya[8].

In terms of pathogen variability, it is well known that the pneumococcal polysaccharide capsule, which determines its serotype, contributes to invasive propensity[1,2]. The pneumococcal genome also encodes a variety of proteins which directly interact with the host, mostly to enhance colonisation and avoid the host immune response[9]. Mouse models have shown that some antigens such *pspC* (*cbpA*) enhance virulence but are not essential in invasive isolates. Though the role of these antigens in colonisation and disease may be known, whether sequence variation at these loci has an effect on pathogenesis in human disease remains unclear. Previous small association studies have additionally suggested a role for platelet binding[10] and arginine synthesis[11] in pneumococcal meningitis, and analysis of within-host variation found that sequence variation of *dlt* and *pde1* are associated with pneumococcal meningitis[12,13].

Thus, because of a lack of large cohort studies combined with detailed clinical metadata, the overall role of pneumococcal variation in clinical cases of meningitis is as yet unknown. It has not been possible to calculate the degree to which different serotypes affect invasiveness compared to other factors (either genetic or environmental), or if there are serotype-independent loci which are involved in invasion. Bacterial genome wide association studies (GWAS) provide a way to identify pneumococcal sequence variation associated with meningitis, independent of genetic background in an unbiased manner. While GWAS is more challenging in bacteria than humans due to strong population structure and high levels of pan-genomic variation, recent methodological advances have helped overcome these issues[14–16].

bioRxiv preprint first posted online Aug. 7, 2018; doi: http://dx.doi.org/10.1101/386078. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

We have collected data and samples from Dutch adults with meningitis between Jan 1, 2006, and July 1, 2014 in the prospective and nationwide MeninGene cohort[4]. Our study addresses previous limitations of genetic studies by collecting a large number of samples of both host and pathogen DNA from culture-proven cases of pneumococcal meningitis, along with detailed clinical metadata (supplementary table 1). We have performed genotyping and whole genome sequencing of this collection, a combined GWAS of host and pathogen in pneumococcal meningitis, the first time such a study has been attempted for a bacterial disease. We performed a GWAS separately in host and pathogen, continuing to develop new approaches to conduct the latter analysis. In both cases we have collected additional cohorts to replicate our findings (figure 1). We also had sequences available from both host and pathogen in 460 cases allowing analysis of interaction effects in a joint GWAS. Using these well characterised cohorts we explain the role of genetic variation of human and pathogen in pneumococcal meningitis.

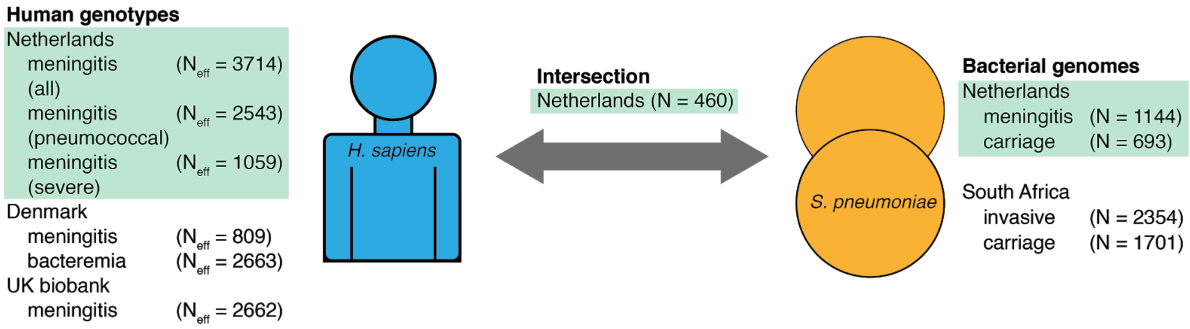

Human genotypes
Netherlands
meningitis (all) (N$_{eff}$ = 3714)
meningitis (pneumococcal) (N$_{eff}$ = 2543)
meningitis (severe) (N$_{eff}$ = 1059)
Denmark
meningitis (N$_{eff}$ = 809)
bacteremia (N$_{eff}$ = 2663)
UK biobank
meningitis (N$_{eff}$ = 2662)
H. sapiens
Intersection
Netherlands (N = 460)
S. pneumoniae
Bacterial genomes
Netherlands
meningitis (N = 1144)
carriage (N = 693)
South Africa
invasive (N = 2354)
carriage (N = 1701)

Figure 1: Overview of cohorts sequenced and associations performed. Left, host data; right, bacterial data; the centre represents samples with both host and pathogen sequence data. Samples in green are those collected from our MeninGene cohort that form the centre of this work. Due to unbalanced case control ratios we show the effective sample size, specific numbers of cases and controls of human genotypes are shown in Supplementary table 2.

bioRxiv preprint first posted online Aug. 7, 2018; doi: http://dx.doi.org/10.1101/386078. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

Results

Human genetics influences pneumococcal meningitis susceptibility and severity

Before trying to find specific loci associated with meningitis, we first calculated the heritability of susceptibility and severity due to host genetics using the MeninGene cohort. While genetic associations have been found with invasive meningococcal disease[7] and susceptibility to infectious diseases more broadly[17,18], the heritability of adult meningitis and its outcome have not previously been calculated. We used two methods to calculate SNP heritability, which showed that variation in host genetics explains 29% ± 7% of the observed variation in pneumococcal meningitis susceptibility, and 49% ± 14% of the variation in meningitis severity (table 1).

Table 1: Human SNP heritability (h^2_{SNP}) of meningitis susceptibility and severity in Dutch adults. Heritabilities are shown on the liability scale (adjusted for population prevalence and case ascertainment ratio). We used two methods for each phenotype, GCTA and LDAK. The latter corrects for linkage disequilibrium when estimating the kinship between genotypes. All results showed significant evidence for a heritability above zero.

Phenotype	Method	Heritability	Error	P-value
Susceptibility (any species)	GCTA	0.22	0.03	2.4×10^{-6}
	LDAK	0.29	0.05	3.9×10^{-11}
Susceptibility (pneumococcal)	GCTA	0.25	0.05	2.4×10^{-6}
	LDAK	0.29	0.07	3.9×10^{-6}
Severity	GCTA	0.29	0.11	2.8×10^{-5}
	LDAK	0.49	0.14	1.4×10^{-4}

Finding that these traits were heritable we then used GWAS to search for specific loci associated with meningitis. Using the MeninGene cohort alone, one marker reached significance when testing for severity: position 64680775 (rs12081070) on chromosome 1, an intronic variant in *UBE2U* which was associated with unfavourable outcome (MAF = 0.43; OR = 1.63; $p = 2.0\times10^{-8}$) (supplementary figures 1-4). *UBE2U* is part of the ubiquitin pathway, but has not been previously associated with any disease or trait. Chromatin conformation capture data shows that the site of the most significantly associated variant interacts with *PGM1* and *ROR1* in a range of immune cell types including monocyte/macrophages, CD4/8 T cells, B cells and neutrophils (supplementary figure 5). *PGM1* encodes a phosphoglucomutase while *ROR1* is a protein of unknown function which has previously been associated with cancers[19] and pulmonary function[20]. There was evidence of association of rs12081070 with gene expression in a panel of tissues and cell

bioRxiv preprint first posted online Aug. 7, 2018; doi: http://dx.doi.org/10.1101/386078. The copyright holder for this preprint
(which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license.

types but this was only significant in skin ($p = 5.7\times10^{-13}$)[21]. Six other loci showed suggestive significance (table 2), whereas in the Danish cohort, no variants reached genome wide significance (supplementary figures 6 & 7).

Table 2: Signals of human association in the MeninGene cohort. We report the lead SNP at each associated locus with MAF > 5% and $p < 1\times10^{-6}$, and nearby annotated genes. The suggestive signal in all meningitis cases was also present when restricted to pneumococcal cases, albeit with a lower p-value of 3.9×10^{-7}.

Position (SNP)	**Effect allele**	**MAF**	**OR**	**P-value**	**Annotation**
Susceptibility (any species)					
chr6:153582990 (rs3870369)	T	0.42	1.27	7.2×10^{-8}	Upstream of *RGS17*
Susceptibility (pneumococcal)					
chr6:117624549 (rs210967)	G	0.46	0.77	8.8×10^{-7}	*ROS1* intronic
chr18:48403560 (rs2850542)	T	0.43	0.65	7.6×10^{-8}	*ME2* promoter (2kb upstream of TSS)
chr22:47506160 (rs13057743)	G	0.33	0.74	5.5×10^{-7}	*TBC1D22A* intronic
Severity					
chr1:64680775 (rs12081070)	A	0.43	1.62	2.0×10^{-8}	*UBE2U* (5th intron)/*ROR1*
chr4:182823804 (rs2309554)	A	0.33	1.58	4.1×10^{-7}	*AC108142.1* intron
chr9:37382231 (rs72739603)	A	0.07	2.36	6.7×10^{-7}	*ZCCHC7/GRHPR*

Of the genes implicated in the single cohort, we noted that the *ME2* promoter variant rs2850542 is an eQTL for the same gene in whole blood ($p = 5.9\times10^{-20}$, supplementary figure 8A) and specifically in monocytes (FDR 5.1x10-26)[22]. There was also evidence of chromatin interaction of the variant location with *SMAD4* again in a range of immune cell types including monocytes, lymphocytes and neutrophils (supplementary figure 8B/C). This gene is involved in TGF-β signalling and invasion across the epithelium; the variant also showed evidence of an eQTL involving rs2850542 for *SMAD4* expression in tibial artery ($p = 8.6\times10^{-7}$)[23].

To improve our discovery power and mitigate false positives from batch effects, we then performed similar associations in the other cohorts. In an analysis of all available cases (MeninGene, Danish invasive disease, GenOSept, 23andme, UK biobank invasive disease) no hits were significantly associated with invasive disease (supplementary figure 9). However, the results for susceptibility to meningitis (MeninGene, Danish meningitis, UK

bioRxiv preprint first posted online Aug. 7, 2018; doi: http://dx.doi.org/10.1101/386078. The copyright holder for this preprint
(which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license.

biobank ICD-10 code for meningitis) found that position 74601544 on chromosome 15 (rs116264669) was associated with the minor allele increasing susceptibility in all three studies ($p = 4.4 \times 10^{-8}$; MAF = 3%) (supplementary figures 10 & 11). This intronic SNP is located in *CDCC33*, a gene that has no prior association with infectious disease. *CDCC33* is expressed in whole blood and the brain[21] although there is no evidence of an eQTL involving this variant or SNPs in linkage disequilibrium with it. The disease-associated variant is located in a genomic region that interacts with the immunoglobulin superfamily containing leucine rich repeat 2 gene *ISLR2* on chromatin conformation capture analysis in macrophages (supplementary figure 12A) and CD8 T-cells (supplementary figure 12B); moreover a variant in complete linkage disequilibrium (rs80140040) with the meningitis susceptibility SNP shows evidence of eQTL with *ISLR2* in a number of tissues including brain ($p = 0.02$). *ISLR2* shows highest expression in the brain (neural tissues, supplementary figure 12C) and plays a role in the development of the nervous system[24] but remains poorly characterised notably in humans. Severity data had not been recorded in other cohorts, and there were too few cases that had resulted in death to allow using this as a proxy for unfavourable outcome in a meta-analysis.

The results from the heritability analysis and GWAS suggest that susceptibility to meningitis is caused by many SNPs with individually low effect sizes (a polygenic trait), and that this cohort size is underpowered to detect these. To determine whether we expected the addition of more samples to uncover new associations with a larger effect size, we used Bayesian mixture models to fit the distribution of effect sizes in the MeninGene cohort. The maximum posterior suggested that 89% of the SNP heritability was caused by 29 large effect size SNPs (oligogenic), with the remaining heritability explained by small effect size SNPs (supplementary table 3). This was further supported by a dynamic Bayesian model incorporating variance in small and large effects, which found support for a mixture of many small effects plus a small number of large effects (maximum posterior $\rho = 0.50$ - proportion of variance explained by sparse terms) (supplementary table 4).

The collection of these datasets gave us the opportunity to perform two further multi-cohort analyses related to sepsis and self-reported meningitis. In the first, we combined Danish bacteremia, GenOSept and self-reported cases of sepsis in the UK biobank, but found no significant hits. Furthermore, we found no evidence that sepsis from clinician diagnosed cases in the UK biobank (using ICD-10 codes) were heritable. Secondly, we meta-analysed 23andme's results for self-reported meningitis with self-reported cases in the UK biobank[17] – the reported hit in *CA10* did not replicate. Given the p-value in the original study was just significant, it is possible that this result was a false positive. This may also be an artefact of respondent's knowledge of whether they had bacterial meningitis, which requires expert knowledge to diagnose and distinguish from viral meningitis[25].

Multiple bacterial loci determine pneumococcal invasive potential

We first performed a heritability analysis to quantify the amount of variation due to the pneumococcal genome for each phenotype on the liability scale. We found that additive pneumococcal genetics explained much of variation in invasive propensity ($h^2 = 85\%$), but no evidence of heritability of meningitis severity ($h^2 = 0\%$). This suggests that invasive

bioRxiv preprint first posted online Aug. 7, 2018; doi: http://dx.doi.org/10.1101/386078. The copyright holder for this preprint
(which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license.

propensity is highly heritable, but that disease outcome is not determined by natural variation of pathogen genetics. The latter is not surprising as invasive disease is an evolutionary dead end for the pathogen so adaptations affecting virulence over the short course of infection are unlikely to be selected for. This is contrary to smaller studies which have suggested that bacterial genotype may help diagnose severe disease in a clinical setting[26], but consistent with a meta-analysis finding no effect of pneumococcal serotype on the risk-ratio of death from meningitis[27].

That pneumococcal invasiveness is affected by pneumococcal genetics is well known, but a quantitative estimate of by how much is unknown. The high heritability estimated here, suggests that in this population some bacteria are able to invade while others are not. This is consistent with some serotypes rarely found in invasive disease[2], and with wide genetic separation from invasive lineages. The current focus of pneumococcal vaccination, and the most well known invasiveness determinant, is serotype. We therefore calculated what proportion of variation in invasiveness can be attributed to the serotype. Although not adjusted directly for genetic background, logistic regression gave a variance in invasiveness explained by the observed serotypes using Nagelkerke's pseudo R^2 from logistic regression of 0.45, less than the total heritability.

Identification of pneumococcal invasiveness loci in multiple cohorts

Given that serotype does not explain all of the variation in invasive potential other pneumococcal factors associated with disease are likely to exist. We then looked at overall differences in the sequence variation between asymptomatic carriage and meningitis isolates. The amount of rare variation compared to common variation present in a population is informative of recent selection and population size changes[28]. An overall difference may therefore be informative of different selection on regions of the genome depending on the niche. In figure 2a we plot the site-frequency spectrum by niche and predicted consequence. Across the range of common minor allele frequencies (MAFs) in both niches the proportion of synonymous/nonsynonymous/intergenic/loss-of-function (LoF) mutations is as previously observed[29]; at low frequencies there is an excess of potentially damaging variants. Figure 2c shows the overall burden of damaging rare variants between carriage and invasive samples; in both LoF and damaging variants there was higher burden in carriage isolates (median LoF: invasive 7, carriage 11, W = 297440, $p < 1 \times 10^{-10}$; median damaging: invasive 22, carriage 26, W = 345370, $p = 8 \times 10^{-4}$),

bioRxiv preprint first posted online Aug. 7, 2018; doi: http://dx.doi.org/10.1101/386078. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

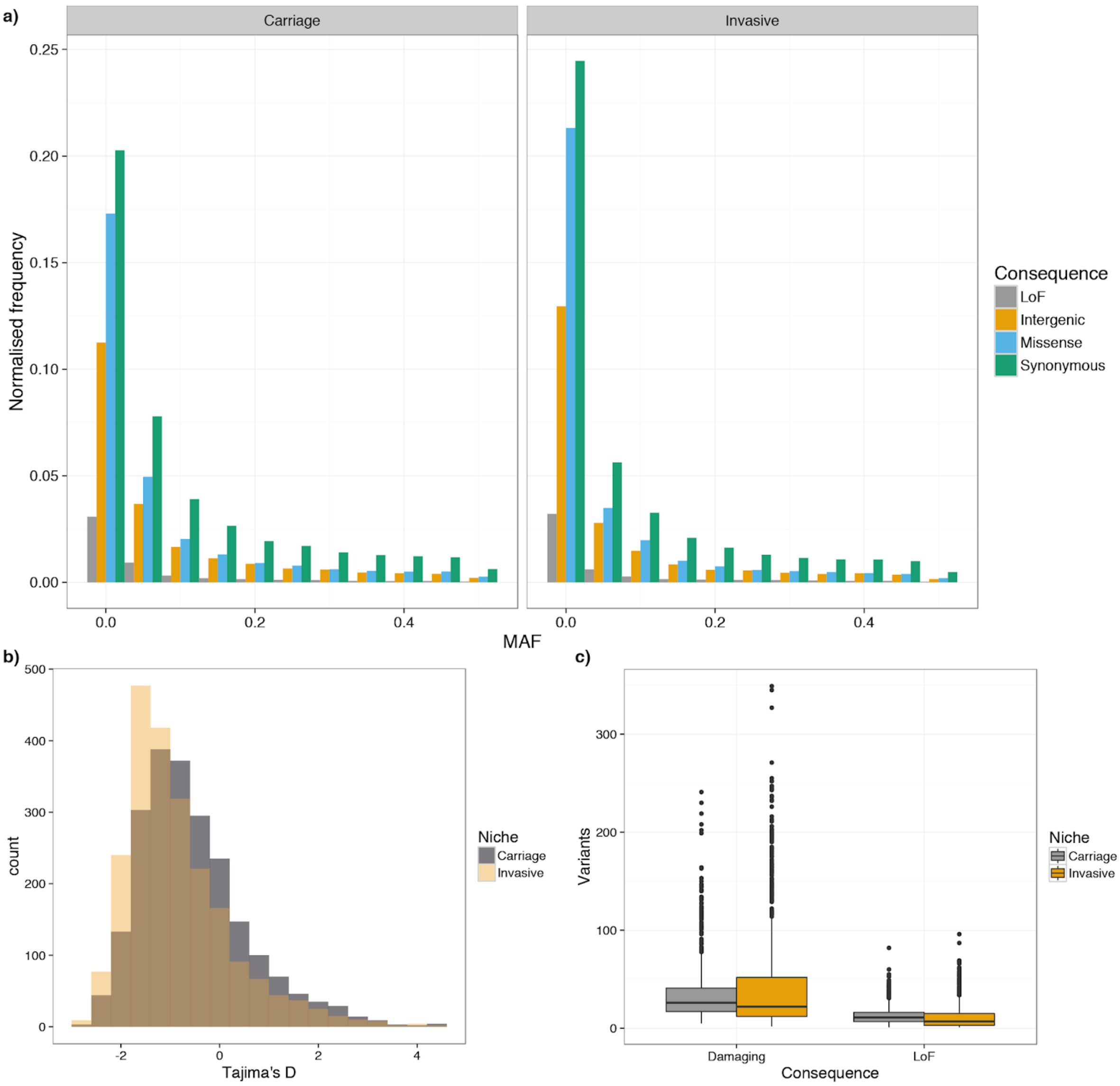


Figure 2: Differing burden and frequency of rare variation between invasive and carriage isolates, based on short variation called from mapping against the ATCC 700669 reference genome. Loss-of-function (LoF) are frameshift or nonsense mutations. a) The site frequency spectrum (SFS) stratified by niche and by predicted consequence. Frequency has been normalised with respect to the number of samples in each population. b) Histogram of Tajima's D for all coding sequences in the genome, stratified by niche. c) Boxplot of number of rare variants per sample, stratified by niche and predicted consequence. Damaging mutations are LoF mutations and missense mutations predicted damaging by PROVEAN.

In total, invasive samples had a higher proportion of rare variants than carriage samples. To quantify this difference and identify which regions of the genome are responsible for the excess of rare alleles we calculated Tajima's D, a statistic for neutral evolution, for each coding sequence in the genome, and looked for differing signs of selection between cases and controls. Deviations with $D < 0$ are indicative of selective sweeps and/or recent population expansion, whereas $D > 0$ is indicative of balancing selection and/or recent population contraction. In terms of differences between site-frequency spectrum (SFS), a negative D manifests as an excess of rare variants whereas a positive D manifests as a uniform distribution. We compared the distributions of D by gene in each phenotype (figure

bioRxiv preprint first posted online Aug. 7, 2018; doi: http://dx.doi.org/10.1101/386078. The copyright holder for this preprint
(which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license.

2b). Genes in invasive isolates had a lower average D (difference in medians -0.34; W = 1996100, $p < 10^{-10}$) and a more positively skewed D (difference in skewness 0.30; 95% bootstrapped CI 0.17-0.44). This difference in D may be representative of a genuine difference in selection of variants in genes between niche, or may be due to a difference in recent population dynamics, for example due to the bottlenecks for invasion and transmission.

We then used GWAS to find locus effects on invasive disease, independent of serotype. As pneumococcal genomes vary extensively in their pan-genome[30], as well as having phenotypically important mosaic structural variants[31] and antigens alleles[32], we used a combination of methods to catalogue the population level variation then tested for associations while adjusting for population structure. We first performed this GWAS with disease severity, which was only measured in the Dutch cohort. Consistent with our estimates of no heritability, we found no loci of any type to be significantly associated with severity.

We then analysed meningitis versus carriage isolates. We first performed this analysis in the Dutch cohort (supplementary tables 5-7) revealing that many of our results involved rare variation, which has largely been ignored in previous bacterial GWAS studies[16]. This variation may be more important in a disease like meningitis where there is no pervasive selection for the phenotype. To improve the power and reliability of our results, we combined our Dutch cohort of meningitis and carriage samples with a cohort collected in South Africa, which included samples from carriage and cases of invasive pneumococcal disease. This gave a total of 5845 pneumococcal genomes to analyse. Table 3 shows the genes which were significant in this combined analysis using any of our association methods.

Table 3: Signals of bacterial association in the combined Dutch and South-African cohorts. Genes significant (Bonferroni corrected $P < 0.05$) in a pooled analysis of both cohorts with any of the association approaches, ordered by p-value. Odds-ratios are with respect to carriage samples. The genes in bold in the top half of the table are immunogenic, and are have previous evidence for association with virulence. For Tajima's D the effect size is the difference between D values, and for k-mers and LoF burden tests it is the odds-ratio. For some p-values the calculation only allows an upper bound to be produced. The locus tag in the ATCC 700669 reference is listed, along with the common gene name if available.

Gene ID	Gene name	Method	AF	OR/Effect size	P-value	Function
SPN23F05680	***ldcB/dacB***	**Missense burden**	**0.81 (gene)**	**1.20**	**1.3×10^{-19}**	**D-alanyl-D-alanine carboxypeptidase (peptoglycan peptide precursors)**
SPN23F22240	***pspC/cbpA***	**K-mers**	**0.2 (k-mer) 0.59 (gene)**	**1.05**	**3.8×10^{-13}**	**Binds secretory IgA, C3 and complement factor H; adhesin**
SPN23F10590	***zmpD***	**Rare LoF in invasive**	**0.53 (gene)**	**1.42**	**$<1 \times 10^{-10}$**	**Unknown; paralogous to IgA1 protease (*zmpA*)**
SPN23F08080	***spnTVRhsdS***	**Missense burden**	**1.00 (gene)**	**1.08**	**2.8×10^{-6}**	**Type I restriction-modification system specificity subunit S**

bioRxiv preprint first posted online Aug. 7, 2018; doi: http://dx.doi.org/10.1101/386078. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

SPN23F17820	***psrP***	**Tajima's D**	**0.42 (gene)**	**-2.71 (invasive)** **-2.59 (carriage)**	**$<1x10^{-6}$**	**adhesin**
SPN23F09820		Missense burden	0.36 (gene)	1.3	$3.4x10^{-49}$	Bacteriocin precursor
SPN23F12140	*ntpI*	Missense burden	0.88 (gene)	0.71	$1.3x10^{-48}$	V-type sodium ATP synthase subunit I
SPN23F05670	*FM211187.1804*	Missense burden	1.00 (gene)	1.27	$4.3x10^{-47}$	Histidine triad family protein (nucleotide phosphatase)
SPN23F04400	*FM211187.1384*	Missense burden	1.00 (gene)	0.76	$6.9x10^{-41}$	Unknown
SPN23F18990	*FM211187.5748*	Missense burden	1.00 (gene)	0.76	$1.9x10^{-18}$	ABC transporter ATP-binding protein
SPN23F01160	*FM211187.339*	Missense burden	1.00 (gene)	0.88	$2.7x10^{-11}$	Unknown
SPN23F20970	*puuD*	Missense burden	1.00 (gene)	1.24	$1.9x10^{-10}$	Gamma-glutamyl-gamma-aminobutyrate hydrolase
SPN23F04740	*ecsA*	Missense burden	1.00 (gene)	1.15	$1.4x10^{-8}$	ABC transporter ATPase
SPN23F16700		Missense burden	1.00 (gene)	1.16	$3.8x10^{-8}$	Nucleotide diphosphate hydrolase
SPN23F21080	*phoR/pnpS*	Missense burden	1.00 (gene)	0.89	$1.1x10^{-6}$	Phosphate-sensitive histidine sensor kinase
SPN23F11460	*mcrB*	Missense burden	-	1.13	$1.2x10^{-6}$	Endonuclease
SPN23F05090		Missense burden	1.00 (gene)	1.15	$1.0x10^{-5}$	Aldose epimerase (putative)
SPN23F12730		Missense burden	-	0.77	$2.1x10^{-5}$	Bacteriocin

The genes noted in bold font in table 3 are all immunogenic[32], and have all previously been associated with pneumococcal virulence in animal models[33–36], but this is the first time an association has been shown in patients with invasive disease. Previous conclusions, drawn from protein binding to the Laminin receptor, have suggested that *pspC* (*cbpA*) is necessary for meningitis[37]. We predicted that both forms of *pspC* were absent in 13 meningitis isolates, though on closer inspection of the summary statistics from mapping and assembly these may also be an unresolved form of allele 8. Using our database of clinical data we also found that all of these patients had a severe ear, nose or throat infection, suggestive for direct spread of bacteria rather than crossing the blood-brain barrier. Three patients had clear bone destruction and/or pneumocephalus, which is proof for direct spread of the infection, and one patient had a skull defect. We further tested whether the two major forms of PspC were associated with meningitis specifically, as has previously been suggested[38], but did not find either to be overrepresented, when accounting for population structure. *dacB* is involved in preserving cell wall shape, and has shown to attenuate virulence in a mouse

bioRxiv preprint first posted online Aug. 7, 2018; doi: http://dx.doi.org/10.1101/386078. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

model of lung infection. *zmpD* is homologous to IgA1 protease (*zmpA*)[39], and while it is immunogenic[32], its function is unknown – these results suggest a role in human cases of disease. The other genes in table 3 have not previously been directly associated with virulence or invasive disease in *S. pneumoniae*.

Interactions between host and pathogen genomes

It is possible that different host genotypes have varying susceptibility to different lineages, or strains carrying certain alleles. To test this we used the 460 samples where we had collected both human genotype and pneumococcal genome sequence (figure 1). In the first instance, to retain hypothesis-free approach of GWAS, we performed a host-pathogen interaction analysis between every pair of common bacterial variants and genotyped host variants. While we were able to perform the $2\text{x}10^{10}$ associations required, no pairs of loci surpassed the large multiple testing burden required by this analysis - a power calculation showed that we would have 80% power for finding an effect with MAF of 25% and OR of 4 (supplementary figure 13) in the absence of population structure. Through this approach, we can therefore rule out large single interaction effects (with OR > 4 and MAF > 25%) between host and pathogen in cases of meningitis.

Given the difficulty of reaching significance for this large number of tests, we then went on to consider regions with strong prior evidence for being involved in host-pathogen interaction. *S. pneumoniae* has many virulence factors, some of which are known to interact with specific human proteins[9]. We were interested in the interactions where the pneumococcal protein contains sequence variation, ideally somewhat independent of genetic background. These regions have a higher power to be detected in an association analysis, and the higher rate of variation is potentially a sign of diversifying selection, which may mean the variation is more likely to be associated with specific interactions with the human immune system. We tested for an association between host genotype and the allele of three antigens selected for their variability and immunogenicity: PspC (CbpA), PspA, and ZmpA. For all of the antigen alleles with enough observations (supplementary table 8) we performed an association against all imputed human variants as above, and using a more accurate imputation of the *CFH* region due to its potential relevance in these interactions.

None of the bacterial antigen alleles were significantly correlated with variants in their human interacting-protein counterparts at the suggestive level ($p < 10^{-5}$). However, there were two associations of a *cbpA* allele reaching genome-wide significance elsewhere in the genome. Supplementary figure 14 shows a locuszoom plot of each of these associations. The first is between *cbpA*-8 and position 148788006 on chromosome 6 (MAF = 0.08; OR = 9.20; p = $4.1\text{x}10^{-9}$). This is in *SASH1*, which has previously been found to have decreased expression during meningococcal meningitis (https://www.ebi.ac.uk/arrayexpress/experiments/E-GEOD-11755/) . The second is between pspC-9 and position 98891272 on chromosome 13 (MAF = 0.16; OR = 6.30; p = $3.6\text{x}10^{-8}$), in *FARP1*, a gene not previously associated with infectious disease. Of note, we could find no published evidence of chromatin confirmation capture interaction or eQTL effects with either of these human variants.

bioRxiv preprint first posted online Aug. 7, 2018; doi: http://dx.doi.org/10.1101/386078. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

We attempted to reduce the multiple testing burden by reducing the dimension of the pathogen genotype, which takes advantage of the extensive genome-wide correlation between variants. To give a straightforward biological interpretation we used hierarchical Bayesian clustering to define lineages, and tested whether pathogen genotypes, so defined, were associated with host genotype. We ran associations with lineages with at least 10% of samples in the subphenotype (supplementary table 9). The only result reaching genome-wide significance was an association between cluster eight (serotypes 9N/15B/19A, which have no overall association with invasive disease over carriage) and variants on chromosome 10 (supplementary figure 15). The lead variant (rs10870273) is at position 134046136 on chromosome 10 (MAF = 0.27; OR = 4.28; $p = 1.2\times10^{-8}$) located in an intron of STK32C, a serine/threonine kinase highly expressed in the brain. The high effect size estimated for the interaction is consistent with the power predicted in supplementary figure 13.

bioRxiv preprint first posted online Aug. 7, 2018; doi: http://dx.doi.org/10.1101/386078. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

Discussion

How genetics can affect susceptibility to and severity of pneumococcal meningitis has not been systematically investigated. We used multiple independent cohorts to perform human and pathogen GWAS to investigate the genetics of pneumococcal meningitis. We found no evidence for pathogen genetics affecting the severity of disease, whereas human genetics explained 49% ± 9% of this variation. In our Dutch cohort, variation near *UBE2U* and *ROR1* was significantly associated with severity. Our findings suggest that sequencing of the pathogen is likely to be uninformative for predicting disease progression, whereas further investigation of the host (and these genes in particular) may lead to greater insights into the mechanism behind severe cases of meningitis.

Host genetics explained 29% ± 7% of variation in susceptibility to meningitis, and a pooled GWAS analysis with our Danish cohort and the UK biobank found an association at *CDCC33*. This gene does not have a currently known function which is related to immunity, so this association may provide a new avenue with which to approach host studies; functional genomic data suggest a further possible mechanism for the susceptibility associated variant through interacting at a distance with and modulating expression of the brain expressed leucine-rich repeat and immunoglobulin (LIG) family protein gene *ISLR2*. By analysing the distribution of effect sizes, we found that meningitis susceptibility is likely to be affected by both a small number of large effect variants (oligogenic), and a large number of small effect variants (polygenic). We found no evidence of association between pneumococcal meningitis and *CFH*, which has previously been associated with meningococcal meningitis[7], nor at any other candidate locus. We were also not able to report a previous finding in *CA10* from self-reported meningitis status, pointing to a need for careful clinical phenotyping needed in studies of hard to diagnose diseases such as bacterial meningitis.

Research of the role of pneumococcal variation in invasive potential in large epidemiological studies has mostly been focused on serotype variation. The lack of cohorts with whole genomes and invasive phenotypes has not allowed determination of other virulence factors in human disease – it is only with large cohorts of whole genome sequences that contributions to pneumococcal phenotypes can be systematically attributed to serotype or other genetic variation. With our large collection of genomes we were able to determine that the bacterial genome is crucial in determining invasive potential, with serotype likely to be the main factor (45% of variance explained). We went on to perform a combined analysis using 5892 pneumococcal genomes from two independent cohorts to find specific variation associated with invasive disease. We found five genes independent of genetic background and serotype to be associated with invasive disease. This showed a role for the virulence genes *cbpA* (*pspC*), *dacB* and *psrP* in human disease for the first time, and a new association with the loss of expression of the immunogenic protein *zmpD*.

Our cohort also allowed a joint analysis of bacterial and human sequencing data. Though the high dimension of data required more samples to find interaction effects of modest effect

bioRxiv preprint first posted online Aug. 7, 2018; doi: http://dx.doi.org/10.1101/386078. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

sizes, through biologically guided dimension reduction we were able to show evidence for possible enrichment of certain pathogen genotypes in certain host genotypes.

Pneumococcal meningitis, although a clinically important area of study due to its poor prognosis, is difficult to study. It is a relatively rare disease, challenging to diagnose in a timely manner, and requires a well set-up study to record the causal pathogen. Severe cases of meningitis are rare and may be difficult to record due to the need to follow up patients. Although there are no more current cohorts available to validate our results in, our findings suggest more cases should be collected to find further associations. Further replication using in vivo models will be necessary to confirm our results associating genes not previously known to be involved in pneumococcal meningitis.

Phenotype heterogeneity, here between causal organism, age of host infected and immunocompromised status may make replication of results with a large number of samples difficult. While we attempted to address this by looking for differences between sub-phenotypes, our sample size was likely too low to draw firm conclusions. It has been proposed that one reason why common variants associated with variation in infectious disease may be hard to find is that over human history they would have been selected against, and are therefore purged from the population[18,40]. Future genetic studies of meningitis should continue to collect samples with detailed clinical data, including the causal pathogen, to minimise phenotype heterogeneity.

bioRxiv preprint first posted online Aug. 7, 2018; doi: http://dx.doi.org/10.1101/386078. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

Methods

Human genotyping and quality control

We performed genotyping using the Illumina Omni array, and called genotypes from normalised intensity data using optiCall[41]. For data taken from other platforms, we merged cases and controls only at sites in the intersection of the genotyping arrays used. We then performed basic quality control steps to first remove low quality samples, then low quality markers[42]. Samples with a heterozygosity rate three standard deviations away from the mean, or over 3% missing genotypes were removed. Markers with over 5% missing genotypes, significantly different ($p < 10^{-5}$) call rate between cases and controls, MAF < 1% or out of Hardy-Weinberg equilibrium ($p < 10^{-5}$) were removed. Using an LD-pruned set of markers, we estimated sample relatedness with KING[43], and removed any duplicate samples. Using the same set of markers, we used eigenstrat to perform a PCA to check sample ancestry (supplementary figure 16)[44]. Samples closer than third-degree relation and samples of non-European ancestry (which we defined as PC1 < 0.07) were removed for heritability analysis, initial association attempts and analysis with Subtest. We manually inspected intensity plots for any associated markers using Evoker[45], and removed any miscalled sites. Finally, we removed markers significantly associated with control batch ($p < 5 \times 10^{-8}$).

All markers were reported with respect to the reference allele and coordinates of GRCh37. We imputed markers using the HRC as a reference panel with the Sanger Imputation Server[46–48]. For the Danish samples, we instead used the Michigan imputation server due to the decreased number of markers available from merging two different genotyping arrays[49]. For greater accuracy, we reimputed the *CFH* region was imputed using impute2 with 1000 Genomes and GoNL as reference panels[50–52]. We removed resulting markers with MAF < 1%, HWE $p < 10^{-5}$ or INFO scores < 0.7 leaving 6.8M markers for association testing and heritability estimation.

Association of human variation

Throughout, we used a Glasgow Outcome Scale score[53] of anything less than five (any long term disability or death) to define unfavourable outcome. We performed the association study using bolt-lmm[54,55], using the LD-pruned set of genotyped markers to estimate the kinship matrix, and then calculating association statistics for all genotyped and imputed sites passing the above quality control thresholds. For the Dutch samples we included whether the patient was immunocompromised as a fixed effect (10% of cases), assuming that no control samples were immunocompromised (1% population prevalence[56,57]). To estimate heritability, we used two methods: GCTA-GREML[58] (as implemented in bolt-lmm) and LDAK v5[59]. For both, we only used samples passing the stricter thresholds for ancestry and relatedness. Estimates of heritability were transformed from the observed scale to the liability scale using a population prevalence of meningitis of 1×10^{-3}. When using Subtest to search for genetic difference between subphenotypes[60], we used default settings, using the weights

bioRxiv preprint first posted online Aug. 7, 2018; doi: http://dx.doi.org/10.1101/386078. The copyright holder for this preprint
(which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license.

per marker calculated by LDAK to adjust for LD, and 1500 draws of 400 subsamples to generate null distributions of the test statistic.

We did not find evidence of overall differences between pneumococcal meningitis and other bacterial meningitis (pseudo-likelihood ratio (PLR) = 0.25; p = 0.75) or between severity and susceptibility (PLR = 0.14; p = 1.00). In the Danish cohort, there was no evidence of difference between meningitis and bacteremia (PLR = 311; p = 0.60). However this technique may rely on relatively highly associated SNPs, which were not found with this few samples. Susceptibility to any meningitis had a significantly higher heritability than its sub-phenotypes, which also have heritability above zero. This is more consistent with some difference in genetic architecture between the phenotypes.

To infer the distributions of effect sizes on meningitis, we ran bayesR[61] and the Bayesian Sparse Linear Mixed Model (BSLMM), as implemented in gemma[62]. For both models, we used the genotyped sites (~630k) in the Dutch data using all meningitis cases as the phenotype. For bayesR we used version 2 of the software, using the '-shuffle' option to increase computational efficiency. We ran the MCMC in each case with default settings: 5×10^5 iterations, discarding the first 2×10^5 as burn-in and sampling every tenth iteration. We ran BSLMM using a probit link function for 1×10^6 iterations, discarding the first 10^5 as burn-in and sampling every tenth iteration. In both cases we report the mean value of each hyperparameter in the posterior.

To perform associations using the UK biobank we used bolt-lmm, following the recommended protocol for analysing the available genetic data[55]. We extracted case samples with self-reported meningitis or sepsis/septicaemia (data-field 20002), with a diagnosis of meningitis (data-field 41202 having a value of G01, G001, G002, G003, G008, A170, A390 or A321 at least once), and with a diagnosis of sepsis (A403, A409, A408 or A40 at least once). We randomly selected 3000 control samples from the remaining samples which had passed the UK biobank genetic quality control, which allowed for quicker analysis with little impact on effective sample size. Using this sample, we removed genotyped markers with MAF < 0.001 or a missing rate > 0.1, and used this to estimate kinships in bolt-lmm. We used bolt-lmm to perform association analysis of every imputed SNP site, including participant age as a fixed effect.

We used METAL to perform meta-analysis between different sets of studies[63]. We used the effective sample size to weight the beta and SE from each set of summary statistics, also adjusting the beta values and standard errors produced by bolt-lmm (supplementary table 10). We only retained markers which had been successfully imputed in all studies, to avoid effects of varying sample size at each locus.

Chromatin conformation capture data was tested and presented using the Capture HiC Plotter[64] and eQTL data from the GTEx Consortium[21].

bioRxiv preprint first posted online Aug. 7, 2018; doi: http://dx.doi.org/10.1101/386078. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

Catalogue of bacterial variation

From the whole genome sequence data of bacteria in the cohort we called SNPs and short INDELs with respect to the ATCC 700669 reference[65]. We mapped reads with bwa mem[66], marked optical duplicates with Picard, and called variants with GATK HaplotypeCaller[67]. For INDELs we used the recommended hard filters. For SNPs we used the recommended hard filters to create an initial call set. We then applied GATK VariantRecalibrator using the following call sites as true positive priors: the intersection of SNPs called by both GATK and bcftools (Q10; 90% confidence); filtered SNPs from a carriage cohort of Karen infants[68] (Q5; 68% confidence); filtered SNPs from a carriage cohort of children in Massachusetts[30] (Q5; 68% confidence). After quality score recalibration we used 99.9% recall as a cutoff for SNPs to maximise sensitivity, and annotated the predicted effect of all coding variation using the variant effect predictor[69]. We defined LoF variants as either stop gained or frameshift mutations. We used PROVEAN with a score cutoff of < -2.5 to predict whether non-synonymous SNPs affected protein function[70].

We produced a core gene alignment using roary[71], where we used reciprocal best BLAST[72] to choose a cut-off of 95% BLAST ID that maximised the balanced accuracy. To produce a presence/absence matrix of accessory elements, to be used as variants in association testing, we mapped the annotated genes in each isolate to a manually curated reference set[32,73]. As a first pass we used cd-hit[74], and then used blastp against the genes which were unclustered after the first pass.

We counted variable length k-mers with a minor allele count of at least ten using fsm-lite[14]. In the Dutch data there were 11.7M informative k-mers, with 2.6M unique patterns. Using the mapping to the ATCC 700669 reference, we used cn.mops[75] to call CNVs. We used windows of 1kb, and used windows with support for a CNV in at least two samples. The number of reads mapping to each *ivr* allele in each sample was determined by using read pair information, as previously[12,31].

While the k-mer approach should directly assay or tag most variation at the population level, the allelic variation of the pneumococcal antigens may not be well captured. For example, *pspC* can be difficult to assemble due to repeats and copy number variation[76], and k-mers from *pspA* and *zmpA* may not map to each allele specifically due to orthologous and paralogous genes[39,77]. As these antigens have been shown to interact with the host immune system, we developed a way to classify the allele present in each isolate combining assembly and mapping statistics. For each antigen in each sample we mapped reads to a reference panel using srst2[78], and aligned annotated genes from the assemblies using blastp. We used coverage, number of SNP mismatches, number of INDEL mismatches and number of truncated bases as summary statistics for each reference allele from srst2, and equivalently percent ID, number of mismatches, number of gaps, E-value and bitscore from blastp. For *pspC*, we used an existing classification scheme of 11 alleles from 48 sequences[76] (supplementary figure 17). For *zmpA* and *pspA* we built trees from previously characterised alleles[32] (supplementary figures 18 & 19). Finding that the ancestral branches in the phylogenies were poorly supported, while the topology of clades were well resolved,

bioRxiv preprint first posted online Aug. 7, 2018; doi: http://dx.doi.org/10.1101/386078. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

we took a cut through the deep branches of the phylogeny to define four allele groups for *pspA* and three for *zmpA*. We ensured the training data were separable into these groups using PCA (supplementary figure 20). We tested the performance of four out-of-the-box classifiers on 20 *pspC* alleles spread across the population that we manually typed from the assemblies (supplementary table 11). Finding that an SVM with a linear kernel worked best, we used this to classify the allele of all antigens in all isolates using the summary statistics described above (supplementary figure 21).

Using the South African samples we counted k-mers, SNPs and INDELs and COGs in the same way, and annotated LoF function variants. We found 52215 SNPs and INDELs, 6.3M informative k-mers, with 1.5M unique patterns.

Association of bacterial variation

Using the SNP and INDEL alignment we built a phylogenetic tree from this alignment using fasttree[79], and calculated the kinship between each pair of strains as the distance between their MRCA and the midpoint root. We calculated heritability on the liability scale[80] using this kinship matrix with FaST-LMM[81]. To estimate contributions for specific variant components such as serotype, we used lasso regression with leave-one-out cross validation to select significant predictors[16] and Nagelkerke's pseudo R^2 to find the variance explained in phenotype[82,83]. To test the effect of capsule charge on phenotype we used previously measured zeta potentials in place of serotype[84], using the serogroup average when a serotype specific charge was not known. Invasiveness was not well predicted from capsule charge alone ($R^2 = 0.08$)[84], partly because of the unknown capsule charge for many of the serotypes observed.

For association of common variation (MAF > 1%) we compared SEER[14], using the first ten multidimensional scaling components as fixed effects to control for population structure, with FaST-LMM[81], which uses eigenvectors from the kinship matrix calculated from the SNP and INDEL alignment as random effects to control for population structure. The Q-Q plots using fixed effects were highly inflated (supplementary figure 22), so we used the linear mixed model throughout. To correct for multiple testing we used the number of unique patterns as the number of tests in a Bonferroni correction, giving $p < 8.2\times10^{-7}$ for SNPs and $p < 1.9\times10^{-8}$ for k-mers. Inspection of the QQ-plots showed inflation of the test statistic for k-mers, so we used a higher threshold of $p < 10^{-16}$ instead. The same association methods was used with the antigen alleles and CNVs.

We considered whether the *ivr* locus, a phase variable inverting type I R-M system with six possible alleles [31], is associated with meningitis. Using mouse models, alleles at this locus have previously been shown to be selected for in carriage, whereas others are preferred in invasive disease[35,85]. The rapid variation of this locus allows simple associations independent of genetic background. To test for association of the *ivr* locus alleles with susceptibility and severity, we used a Bayesian hierarchical model we had previously developed to find differences in the proportion of alleles present in tissue types while accounting for heterogeneity within single colony picks[12]. We used the same priors (using allele prevalence calculated using long-range PCR from a subset of samples) and MCMC parameters as

bioRxiv preprint first posted online Aug. 7, 2018; doi: http://dx.doi.org/10.1101/386078. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

specified previously, but labelling with phenotype rather than tissue. We found no evidence that either allele frequencies or overall diversity had any association with invasive disease or carriage in clinical cases of meningitis (supplementary figures 23 & 24).

To calculate Tajima's D between phenotype groups we wrote a program in C++ using a function to calculate pairwise differences between strains we had previously optimised. Gaps or unknown sites were ignored. We calculated *D* for all coding sequences annotated in the ATCC 700669 reference, and p-values for difference between niche were calculated using 44000 null permutations of phenotype labels[86]. We applied a Bonferroni correction using the number of coding sequences tested.

Rare variants (MAF < 1%) could not be directly associated. Instead we applied a burden test[87], grouping variants by coding sequence. As burden tests lose power when variants have different directions of effect on the phenotype, we used only those variants predicted to cause a loss of function in one test, and those causing either loss of function (6825 variants) or predicted change in protein function in another (additional 26206 variants). These variants will have occurred on terminal (or close to terminal) branches and therefore population structure is less of an issue than for common variants. We used plink/seq to perform this association for each phenotype, applying a Bonferroni correction using the number of genes as the number of multiple tests.

Reasoning that power to detect individual variants may also be hampered by population structure as well as allele frequency, we also searched for a burden of missense variants of any frequency by gene. We used the LMM burden testing mode of pyseer[88] which allowed us to correct for population structure with the same model as above.

When performing analysis with both the Dutch and South African samples we pooled the genetic data, and performed the same association analysis as for the Dutch data alone. Where possible (associations using the linear mixed model) we included country as a covariate. The South African cohort also included host gender, age, collection year and HIV status and PCV-use at time of sampling. We included these as additional covariates for these samples. We were not able to use PROVEAN with this data, so only performed a burden test of damaging LoF variants. We used a significance threshold of $P < 0.05$ in the pooled analysis, after applying a Bonferroni correction for multiple testing based on the number of unique patterns as above. For all tests we ensured that the QQ-plots of the resulting p-values were not inflated (supplementary figure 25). We observed hits to simple transposons without cargo genes, which we therefore discarded as we assumed this was an artefact of their independence from population structure. We also found significant association of some BOX repeats[89], but as there are many copies we could not map these associations to a single region of the genome.

Interaction effects

We took 460 pneumococcal meningitis samples with matched pathogen and human sequence data which passed quality control thresholds for both data types. This has previously been applied to coding changes and host genotypes for HIV[90] and HCV[91], though

bioRxiv preprint first posted online Aug. 7, 2018; doi: http://dx.doi.org/10.1101/386078. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

these have much less variation than the pneumococcal population. To test all variants in a pairwise manner we converted the VCFs of human and pathogen calls into CSV files treating human genotypes as additive, and storing site and sample data separately for more efficient access by chunk[92]. The number of pairwise tests between all common variants was prohibitively large (10^{12}), so we only tested genotyped markers: 1.8×10^{10} pairs of variants passed filters of MAF > 5% and missing rate < 5% in both the human and pathogen data. We modified the association code of SEER [14] to extend the χ^2 test to a 3x2 table, and to perform a 3x2 Fisher's exact test (using https://github.com/chrchang/stats) when assumptions of the χ^2 test were violated. Those sites with $p < 5 \times 10^{-11}$ (a Bonferroni correction with alpha = 1, as an initial filter) were then tested using a logistic regression of the human SNP against the pathogen variant, with the first three components from multidimensional scaling of the pathogen kinship matrix included as covariates to adjust for pathogen population structure.

To test for an association between invading lineage and human genotype we re-ran hierBAPS on the 460 samples[93], which generated ten top level clusters (including a bin cluster) seven of which were large enough to test (supplementary table 9). We used bolt-lmm as above, but used whether the invading genotype was a member of the BAPS cluster as cases. We tested association of antigen alleles with frequencies over 10% in the sampled population in the same way (supplementary table 8).

bioRxiv preprint first posted online Aug. 7, 2018; doi: http://dx.doi.org/10.1101/386078. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

Acknowledgements

We would like to thank Chao Tan and David Hinds from 23andme for sharing summary statistics from their association with self-reported bacterial meningitis. Thanks to Win Kit Man for performing DNA isolation of bacteria. We also thank all members of the GenOSept consortium for making their data available and Anna Rautanen in particular for her preparation of the genetic data. This study makes use of data generated by the Genome of the Netherlands Project (GoNL). Funding for the GoNL project was provided by the Netherlands Organization for Scientific Research under award number 184021007, dated July 9, 2009 and made available as a Rainbow Project of the Biobanking and Biomolecular Research Infrastructure Netherlands (BBMRI-NL). Samples where contributed by LifeLines (http://lifelines.nl/lifelines-research/general), The Leiden Longevity Study (http://www.healthy-ageing.nl; http://www.langleven.net), The Netherlands Twin Registry (NTR: http://www.tweelingenregister. org), The Rotterdam studies, (http://www.erasmus-epidemiology.nl/rotterdamstudy) and the Genetic Research in Isolated Populations program (http://www.epib.nl/research/geneticepi/research.html#gip). The sequencing was carried out in collaboration with the Beijing Institute for Genomics (BGI).

Funding

Work at the Wellcome Trust Sanger Institute was supported by Wellcome Trust (098051). This work was also supported by grants from the European Research Council (ERC Starting Grant, proposal/contract 281156; https://erc.europa.eu/) to DvdB and the Netherlands Organization for Health Research and Development (ZonMw; NWO-Vidi grant 2010, proposal/contract 016.116.358 to DvdB; NWO-Veni Grant NWO-Veni grant 2012 proposal/contract 916.13.078 to MB; www.zonmw.nl/). The Netherlands Reference Laboratory for Bacterial Meningitis was supported by the National Institute for Health and Environmental Protection, Bilthoven (www.rivm.nl/). JAL was supported by a Medical Research Council studentship grant (1365620). NJC was supported by a Sir Henry Dale Fellowship, jointly funded by the Wellcome Trust and the Royal Society (grant number 104169/Z/14/Z). JNW is funded by grants from the United States Public Health Service (AI038446 and AI105168). The BPROOF cohort was funded by The Netherlands Organization for Health Research and Development (ZonMw, Grant 6130.0031), the Hague; unrestricted grant from NZO (Dutch Dairy Association), Zoetermeer; Orthica, Almere; NCHA (Netherlands Consortium Healthy Ageing) Leiden/Rotterdam; Ministry of Economic Affairs, Agriculture and Innovation (project KB-15-004-003), the Hague; Wageningen University, Wageningen; VUmc, Amsterdam; Erasmus Medical Center, Rotterdam; Unilever, Colworth, UK. Dutch Carriage studies at RIVM were funded by the Dutch Ministry of Health, Welfare and Sport. Genotyping for GOYA was funded by the Wellcome Trust (WT 084762MA). GOYA is a nested study within The Danish National Birth Cohort which was established with major funding from the Danish National Research Foundation. Additional support for this cohort has been obtained from the Pharmacy Foundation, the Egmont Foundation, The March of Dimes Birth Defects Foundation, the Augustinus Foundation, and the Health Foundation. Funding for the Danish Invasive Pneumococcal Disease Cohort was kindly provided by the Lundbeck Foundation, the Novo Nordisk Foundation, King Christian the 10th Foundation, Jacob Madsen's Foundation,

bioRxiv preprint first posted online Aug. 7, 2018; doi: http://dx.doi.org/10.1101/386078. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

TrygVesta, Ebba Celinder's Foundation, the Danish Medical Association Foundation, the Foundation for Advancement of Medical Science, the Augustinus Foundation, Peder Laurits Pedersen's Foundation, A.P. Møller Foundation for the Advancement of Medical Science, the Danish Medical Research Council, Preben and Anna Simonsen's Foundation, Ferdinand and Ellen Hindsgaul's Foundation, the Hartmann's Foundation and Dagmar Marshalls Fond. The GenOSept study was supported by the European Union (6th framework programme of RTD funding) and the National Institute for Health Research, through the Comprehensive Clinical Research Network. JCK was supported by NIHR Oxford Biomedical Research Centre and a Wellcome Trust Investigator Award (204969/Z/16/Z) and AJM was supported by a Wellcome Trust Fellowship with reference 106289/Z/14/Z. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author contributions

Conceptualization: JAL, BF, PHCK, JP, AvG, AvdE, MCB, JCB, SDB, DvdB. Data curation: JAL, BF, PHCK, NEW, NJC, RAG, HJB, ZBH, LHÄ, AJM, TCM. Formal analysis: JAL, BF, PHCK, NEW, AHZ, AJM, JCK. Funding acquisition: JP, SM, TB, AvG, AvdE, MCB, JCB, SDB, DvdB. Investigation: JAL, BF, PHCK, NEW, MVS, AJM, TCM, JCK. Methodology: JAL, BF, PHCK, NEW. Project administration: AvdE, MCB, JCB, SDB, DvdB. Resources: MVS, HJB, NR, AJWM, EAMS, KT, ALW, LHvdB, WvR, JHV, ZBH, LFL, LCPGMdG, NMvS, LHÄ, TIAS, EAN, MdP. Software: JAL. Supervision: JNW, JP, SM, TB, AvG, AvdE, MCB, JCB, SDB, DvdB. Writing – original draft: JAL, SDB, DvdB. Writing – review & editing: all authors.

Data and code availability

Code used for the analysis, along with phylogenies and predicted antigen alleles can be found at https://github.com/johnlees/meningene. Bacterial metadata, including ENA accession numbers, can be accessed at on figshare (doi:10.6084/m9.figshare.5915314).

Conflict of interest statement

NJC and SDB were consultants for Antigen Discovery, Inc involved in the design of a proteome array for *S. pneumoniae*. EAMS reports grants from the pharmaceutical companies GlaxoSmithKline and Pfizer outside the submitted work. KT reports grants from Pfizer and consultancy fees from Pfizer paid to University Medical Centre Utrecht, both received outside the submitted work. DvdB received departmental honoraria for serving on a scientific advisory board for GlaxoSmithKline paid to the Academic Medical Center, outside the submitted work.

bioRxiv preprint first posted online Aug. 7, 2018; doi: http://dx.doi.org/10.1101/386078. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

Ethics statement

For the MeninGene study written informed consent was obtained from all patients or their legally authorized representatives. The studies were approved by the Medical Ethics Committee of the Academic Medical Center, Amsterdam, The Netherlands (approval number: NL43784.018.13). For bacterial carriage samples from the Netherlands, written informed consent was obtained from both parents of each child participant, and from all adult participants. Approvals for the 2009 and 2012/2013 studies in children and their parents (NL24116 and NL40288/NTR3614) and for the study in elderly adults (NTR3386) were received from a National Ethics Committee in the Netherlands (CCMO and METC Noord-Holland). For the 2010/2011 study a National Ethics Committee in the Netherlands (the STEG-METC, Almere) decided that approval was not necessary. The studies were conducted in accordance with the European Statements for Good Clinical Practice and the Declaration of Helsinki of the World Medical Association. The Danish Invasive Pneumococcal Disease Cohort was approved by Danish Data Protection Agency (record no. 2007-41-0229 and 01864 HVH-2012-046). Ethical permission was obtained from The Ethical Committee of The Capital Region of Denmark (H-B-2007-085 and H-1-2012-063). According to Danish Legislation, the Research Ethics Committee can grant an exemption from obtaining informed consent for research projects based on biological material under certain circumstances, and for this study such an exemption was granted.

References

1. Brueggemann, A. B. *et al.* Clonal relationships between invasive and carriage Streptococcus pneumoniae and serotype- and clone-specific differences in invasive disease potential. *J. Infect. Dis.* **187,** 1424–1432 (2003).
2. Hausdorff, W. P., Feikin, D. R. & Klugman, K. P. Epidemiological differences among pneumococcal serotypes. *Lancet Infect. Dis.* **5,** 83–93 (2005).
3. Mook-Kanamori, B. B., Geldhoff, M., van der Poll, T. & van de Beek, D. Pathogenesis and pathophysiology of pneumococcal meningitis. *Clin. Microbiol. Rev.* **24,** 557–591 (2011).
4. Bijlsma, M. W. *et al.* Community-acquired bacterial meningitis in adults in the Netherlands, 2006–14: a prospective cohort study. *Lancet Infect. Dis.* **16,** 339–347 (2016/3).
5. Brouwer, M. C. *et al.* Nationwide implementation of adjunctive dexamethasone therapy

bioRxiv preprint first posted online Aug. 7, 2018; doi: http://dx.doi.org/10.1101/386078. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

for pneumococcal meningitis. *Neurology* **75,** 1533–1539 (2010).

6. Brouwer, M. C. *et al.* Host genetic susceptibility to pneumococcal and meningococcal disease: a systematic review and meta-analysis. *Lancet Infect. Dis.* **9,** 31–44 (2009).
7. Davila, S. *et al.* Genome-wide association study identifies variants in the CFH region associated with host susceptibility to meningococcal disease. *Nat. Genet.* **42,** 772–776 (2010).
8. Rautanen, A. *et al.* Polymorphism in a lincRNA Associates with a Doubled Risk of Pneumococcal Bacteremia in Kenyan Children. *Am. J. Hum. Genet.* **2,** 1092–1100 (2016).
9. Kadioglu, A., Weiser, J. N., Paton, J. C. & Andrew, P. W. The role of Streptococcus pneumoniae virulence factors in host respiratory colonization and disease. *Nat. Rev. Microbiol.* **6,** 288–301 (2008).
10. Tunjungputri, R. N. *et al.* Phage-Derived Protein Induces Increased Platelet Activation and Is Associated with Mortality in Patients with Invasive Pneumococcal Disease. *MBio* **8,** (2017).
11. Piet, J. R. *et al.* Streptococcus pneumoniae arginine synthesis genes promote growth and virulence in pneumococcal meningitis. *J. Infect. Dis.* **209,** 1781–1791 (2014).
12. Lees, J. A. *et al.* Large scale genomic analysis shows no evidence for pathogen adaptation between the blood and cerebrospinal fluid niches during bacterial meningitis. *Microb Genom* **3,** e000103 (2017).
13. Lees, J. A. *et al.* Within-Host Sampling of a Natural Population Shows Signs of Selection on Pde1 during Bacterial Meningitis. *Infect. Immun.* **85,** (2017).
14. Lees, J. A. *et al.* Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes. *Nat. Commun.* **7,** 12797 (2016).
15. Earle, S. G. *et al.* Identifying lineage effects when controlling for population structure improves power in bacterial association studies. *Nature Microbiology* 16041 (2016).

bioRxiv preprint first posted online Aug. 7, 2018; doi: http://dx.doi.org/10.1101/386078. The copyright holder for this preprint
(which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license.

16. Lees, J. A. *et al.* Genome-wide identification of lineage and locus specific variation associated with pneumococcal carriage duration. *Elife* **6,** (2017).
17. Tian, C. *et al.* Genome-wide association and HLA region fine-mapping studies identify susceptibility loci for multiple common infections. *Nat. Commun.* **8,** 599 (2017).
18. Chapman, S. J. & Hill, A. V. S. Human genetic susceptibility to infectious disease. *Nat. Rev. Genet.* **13,** 175–188 (2012).
19. Reddy, U. R., Phatak, S. & Pleasure, D. Human neural tissues express a truncated Ror1 receptor tyrosine kinase, lacking both extracellular and transmembrane domains. *Oncogene* **13,** 1555–1559 (1996).
20. Lutz, S. M. *et al.* A genome-wide association study identifies risk loci for spirometric measures among smokers of European and African ancestry. *BMC Genet.* **16,** 138 (2015).
21. GTEx Consortium *et al.* Genetic effects on gene expression across human tissues. *Nature* **550,** 204–213 (2017).
22. Fairfax, B. P. *et al.* Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science* **343,** 1246949 (2014).
23. Clarke, T. B., Francella, N., Huegel, A. & Weiser, J. N. Invasive bacterial pathogens exploit TLR-mediated downregulation of tight junction components to facilitate translocation across the epithelium. *Cell Host Microbe* **9,** 404–414 (2011).
24. Mandai, K., Reimert, D. V. & Ginty, D. D. Linx mediates interaxonal interactions and formation of the internal capsule. *Neuron* **83,** 93–103 (2014).
25. Khatib, U., van de Beek, D., Lees, J. A. & Brouwer, M. C. Adults with suspected central nervous system infection: A prospective study of diagnostic accuracy. *J. Infect.* (2016). doi:10.1016/j.jinf.2016.09.007
26. Cremers, A. J. H. *et al.* The contribution of genetic variation of Streptococcus pneumoniae to the clinical manifestation of invasive pneumococcal disease. *bioRxiv*

bioRxiv preprint first posted online Aug. 7, 2018; doi: http://dx.doi.org/10.1101/386078. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

169722 (2017). doi:10.1101/169722

27. Weinberger, D. M. *et al.* Association of serotype with risk of death due to pneumococcal pneumonia: a meta-analysis. *Clin. Infect. Dis.* **51,** 692–699 (2010).
28. Yang, Z. *Computational Molecular Evolution*. (OUP Oxford, 2006).
29. Thorpe, H. A., Bayliss, S. C., Hurst, L. D. & Feil, E. J. Comparative Analyses of Selection Operating on Nontranslated Intergenic Regions of Diverse Bacterial Species. *Genetics* **206,** 363–376 (2017).
30. Croucher, N. J. *et al.* Population genomics of post-vaccine changes in pneumococcal epidemiology. *Nat. Genet.* **45,** 656–663 (2013).
31. Croucher, N. J. *et al.* Diversification of bacterial genome content through distinct mechanisms over different timescales. *Nat. Commun.* **5,** 5471 (2014).
32. Croucher, N. J. *et al.* Diverse evolutionary patterns of pneumococcal antigens identified by pangenome-wide immunological screening. *Proc. Natl. Acad. Sci. U. S. A.* (2017). doi:10.1073/pnas.1613937114
33. Ogunniyi, A. D. *et al.* Contributions of pneumolysin, pneumococcal surface protein A (PspA), and PspC to pathogenicity of Streptococcus pneumoniae D39 in a mouse model. *Infect. Immun.* **75,** 1843–1851 (2007).
34. Abdullah, M. R. *et al.* Structure of the pneumococcal l,d-carboxypeptidase DacB and pathophysiological effects of disabled cell wall hydrolases DacA and DacB. *Mol. Microbiol.* **93,** 1183–1206 (2014).
35. Manso, A. S. *et al.* A random six-phase switch regulates pneumococcal virulence via global epigenetic changes. *Nat. Commun.* **5,** 5055 (2014).
36. Shivshankar, P., Sanchez, C., Rose, L. F. & Orihuela, C. J. The Streptococcus pneumoniae adhesin PsrP binds to Keratin 10 on lung cells. *Mol. Microbiol.* **73,** 663–679 (2009).
37. Orihuela, C. J. *et al.* Laminin receptor initiates bacterial contact with the blood brain

bioRxiv preprint first posted online Aug. 7, 2018; doi: http://dx.doi.org/10.1101/386078. The copyright holder for this preprint
(which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-ND 4.0 International license.

barrier in experimental meningitis models. *J. Clin. Invest.* **119,** 1638–1646 (2009).

38. van der Maten, E. *et al.* Streptococcus pneumoniae PspC Subgroup Prevalence in Invasive Disease and Differences in Contribution to Complement Evasion. *Infect. Immun.* **86,** (2018).
39. Bek-Thomsen, M., Poulsen, K. & Kilian, M. Occurrence and evolution of the paralogous zinc metalloproteases IgA1 protease, ZmpB, ZmpC, and ZmpD in Streptococcus pneumoniae and related commensal species. *MBio* **3,** (2012).
40. Casanova, J.-L. Severe infectious diseases of childhood as monogenic inborn errors of immunity. *Proc. Natl. Acad. Sci. U. S. A.* **112,** E7128–37 (2015).
41. Shah, T. S. *et al.* optiCall: a robust genotype-calling algorithm for rare, low-frequency and common variants. *Bioinformatics* **28,** 1598–1603 (2012).
42. Anderson, C. a. *et al.* Data quality control in genetic case-control association studies. *Nat. Protoc.* **5,** 1564–1573 (2010).
43. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26,** 2867–2873 (2010).
44. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38,** 904–909 (2006).
45. Morris, J. a., Randall, J. C., Maller, J. B. & Barrett, J. C. Evoker: a visualization tool for genotype intensity data. *Bioinformatics* **26,** 1786–1787 (2010).
46. Loh, P.-R. *et al.* Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet.* **48,** 1443–1448 (2016).
47. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* (2016). doi:10.1038/ng.3643
48. Durbin, R. Efficient haplotype matching and storage using the positional Burrows-Wheeler transform (PBWT). *Bioinformatics* **30,** 1266–1272 (2014).
49. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat. Genet.*

bioRxiv preprint first posted online Aug. 7, 2018; doi: http://dx.doi.org/10.1101/386078. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

48, 1284–1287 (2016).

50. Howie, B., Marchini, J. & Stephens, M. Genotype imputation with thousands of genomes. *G3* **1,** 457–470 (2011).
51. 1000 Genomes Project Consortium *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).
52. The Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat. Genet.* **46,** 818–825 (2014).
53. Jennett, B. & Bond, M. Assessment of outcome after severe brain damage. *Lancet* **1,** 480–484 (1975).
54. Loh, P.-R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47,** 284–290 (2015).
55. Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A. P. & Price, A. L. Mixed model association for biobank-scale data sets. *bioRxiv* 194944 (2017). doi:10.1101/194944
56. van Veen, M. G. *et al.* National estimate of HIV prevalence in the Netherlands: comparison and applicability of different estimation tools. *AIDS* **25,** 229–237 (2011).
57. Harpaz, R., Dahl, R. & Dooling, K. The Prevalence of Immunocompromised Adults: United States, 2013. *Open Forum Infect Dis* **3,** (2016).
58. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88,** 76–82 (2011).
59. Speed, D., Hemani, G., Johnson, M. R. & Balding, D. J. Improved heritability estimation from genome-wide SNPs. *Am. J. Hum. Genet.* **91,** 1011–1021 (2012).
60. Liley, J., Todd, J. A. & Wallace, C. A method for identifying genetic heterogeneity within phenotypically defined disease subgroups. *Nat. Genet.* **49,** 310–316 (2017).
61. Moser, G. *et al.* Simultaneous discovery, estimation and prediction analysis of complex traits using a bayesian mixture model. *PLoS Genet.* **11,** e1004969 (2015).

bioRxiv preprint first posted online Aug. 7, 2018; doi: http://dx.doi.org/10.1101/386078. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

62. Zhou, X., Carbonetto, P. & Stephens, M. Polygenic modeling with bayesian sparse linear mixed models. *PLoS Genet.* **9,** e1003264 (2013).

63. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26,** 2190–2191 (2010).

64. Javierre, B. M. *et al.* Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell* **167,** 1369–1384.e19 (2016).

65. Croucher, N. J. *et al.* Role of conjugative elements in the evolution of the multidrug-resistant pandemic clone Streptococcus pneumoniaeSpain23F ST81. *J. Bacteriol.* **191,** 1480–1489 (2009).

66. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 3 (2013).

67. Van der Auwera, G. A. *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* **43,** 11.10.1–33 (2013).

68. Chewapreecha, C. *et al.* Dense genomic sampling identifies highways of pneumococcal recombination. *Nat. Genet.* **46,** 305–309 (2014).

69. McLaren, W. *et al.* Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26,** 2069–2070 (2010).

70. Choi, Y., Sims, G. E., Murphy, S., Miller, J. R. & Chan, A. P. Predicting the functional effect of amino acid substitutions and indels. *PLoS One* **7,** e46688 (2012).

71. Page, A. J. *et al.* Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31,** btv421 (2015).

72. Ward, N. & Moreno-Hagelsieb, G. Quickly finding orthologs as reciprocal best hits with BLAT, LAST, and UBLAST: how much do we miss? *PLoS One* **9,** e101850 (2014).

73. Corander, J. *et al.* Frequency-dependent selection in vaccine-associated pneumococcal

bioRxiv preprint first posted online Aug. 7, 2018; doi: http://dx.doi.org/10.1101/386078. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

population dynamics. *Nature Ecology & Evolution* 1 (2017).

74. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28,** 3150–3152 (2012).
75. Klambauer, G. *et al.* cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic Acids Res.* **40,** e69 (2012).
76. Iannelli, F., Oggioni, M. R. & Pozzi, G. Allelic variation in the highly polymorphic locus pspC of Streptococcus pneumoniae. *Gene* **284,** 63–71 (2002).
77. Hollingshead, S. K., Becker, R. & Briles, D. E. Diversity of PspA: Mosaic genes and evidence for past recombination in Streptococcus pneumoniae. *Infect. Immun.* **68,** 5889–5900 (2000).
78. Inouye, M. *et al.* SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. *Genome Med.* **6,** 90 (2014).
79. Price, M. N., Dehal, P. S. & Arkin, A. P. Fasttree: Computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.* **26,** 1641–1650 (2009).
80. Lynch, M. & Walsh, B. *Genetics and Analysis of Quantitative Traits*. (Sinauer, 1998).
81. Lippert, C. *et al.* FaST linear mixed models for genome-wide association studies. *Nat. Methods* **8,** 833–835 (2011).
82. International Schizophrenia Consortium *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460,** 748–752 (2009).
83. Hosmer, D. W., Jr., Lemeshow, S. & Sturdivant, R. X. *Applied Logistic Regression*. (John Wiley & Sons, 2013).
84. Li, Y., Weinberger, D. M., Thompson, C. M., Trzciński, K. & Lipsitch, M. Surface charge of Streptococcus pneumoniae predicts serotype distribution. *Infect. Immun.* **81,** 4519–4524 (2013).
85. Li, J. *et al.* Epigenetic Switch Driven by DNA Inversions Dictates Phase Variation in

bioRxiv preprint first posted online Aug. 7, 2018; doi: http://dx.doi.org/10.1101/386078. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

Streptococcus pneumoniae. *PLoS Pathog.* **12,** e1005762 (2016).

86. Winantea, J. *et al.* A summary statistic approach to sequence variation in noncoding regions of six schizophrenia-associated gene loci. *Eur. J. Hum. Genet.* **14,** 1037–1043 (2006).
87. Li, B. & Leal, S. M. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet.* **83,** 311–321 (2008).
88. Lees, J. A., Galardini, M., Bentley, S. D., Weiser, J. N. & Corander, J. pyseer: a comprehensive tool for microbial pangenome-wide association studies. *Bioinformatics* (2018). doi:10.1093/bioinformatics/bty539
89. Croucher, N. J., Vernikos, G. S., Parkhill, J. & Bentley, S. D. Identification, variation and transcription of pneumococcal repeat sequences. *BMC Genomics* **12,** 120 (2011).
90. Bartha, I. *et al.* A genome-to-genome analysis of associations between human genetic variation, HIV-1 sequence diversity, and viral control. *Elife* **2,** 1–16 (2013).
91. Azim Ansari, M. *et al.* Genome-to-genome analysis highlights the effect of the human innate and adaptive immune systems on the hepatitis C virus. *Nat. Genet.* (2017). doi:10.1038/ng.3835
92. Ganna, A. *et al.* Ultra-rare disruptive and damaging mutations influence educational attainment in the general population. *Nat. Neurosci.* **19,** 1563–1565 (2016).
93. Cheng, L., Connor, T. R., Sirén, J., Aanensen, D. M. & Corander, J. Hierarchical and spatially explicit clustering of DNA sequences with BAPS software. *Mol. Biol. Evol.* **30,** 1224–1228 (2013).

The NEW ENGLAND JOURNAL of MEDICINE

REVIEW ARTICLE

DRUG THERAPY

Alastair J.J. Wood, M.D., *Editor*

Pharmacogenomics — Drug Disposition, Drug Targets, and Side Effects

William E. Evans, Pharm.D., and Howard L. McLeod, Pharm.D.

From St. Jude Children's Research Hospital and the University of Tennessee Colleges of Pharmacy and Medicine, Memphis (W.E.E.); and Washington University Medical School, St. Louis (H.L.M.). Address reprint requests to Dr. Evans at St. Jude Children's Research Hospital, 332 N. Lauderdale St., Memphis, TN 38101-0318, or at william.evans@stjude.org.

IT IS WELL RECOGNIZED THAT DIFFERENT PATIENTS RESPOND IN DIFFERent ways to the same medication. These differences are often greater among members of a population than they are within the same person at different times (or between monozygotic twins).[1] The existence of large population differences with small intrapatient variability is consistent with inheritance as a determinant of drug response; it is estimated that genetics can account for 20 to 95 percent of variability in drug disposition and effects.[2] Although many nongenetic factors influence the effects of medications, including age, organ function, concomitant therapy, drug interactions, and the nature of the disease, there are now numerous examples of cases in which interindividual differences in drug response are due to sequence variants in genes encoding drug-metabolizing enzymes, drug transporters, or drug targets.[3-5] Unlike other factors influencing drug response, inherited determinants generally remain stable throughout a person's lifetime.

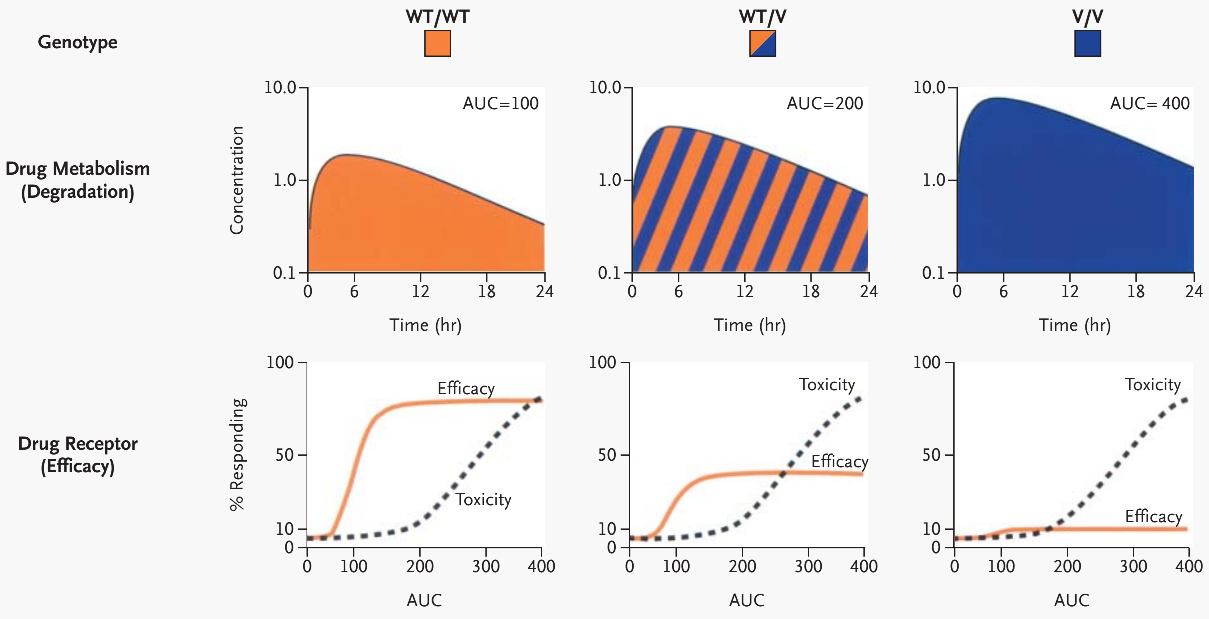
Clinical observations of inherited differences in drug effects were first documented in the 1950s,[6-9] giving rise to the field of pharmacogenetics, and later pharmacogenomics. Although the two terms are synonymous for all practical purposes, pharmacogenomics uses genome-wide approaches to elucidate the inherited basis of differences between persons in the response to drugs.

More than 1.4 million single-nucleotide polymorphisms were identified in the initial sequencing of the human genome,[10] with over 60,000 of them in the coding region of genes. Some of these single-nucleotide polymorphisms have already been associated with substantial changes in the metabolism or effects of medications, and some are now being used to predict clinical response.[3-5,11] Because most drug effects are determined by the interplay of several gene products that influence the pharmacokinetics and pharmacodynamics of medications, including inherited differences in drug targets (e.g., receptors) and drug disposition (e.g., metabolizing enzymes and transporters), polygenic determinants of drug effects (Fig. 1) have become increasingly important in pharmacogenomics. In this review, we focus on the therapeutic consequences of inherited differences in drug disposition and drug targets. An accompanying review[12] focuses on the pharmacogenetics of drug metabolism. This review is not meant to be exhaustive; rather, clinically relevant examples are used to illustrate how pharmacogenomics can provide molecular diagnostic methods that improve drug therapy.

GENETIC POLYMORPHISMS INFLUENCING DRUG DISPOSITION

The field of pharmacogenetics began with a focus on drug metabolism,[12] but it has been extended to encompass the full spectrum of drug disposition, including a growing list of transporters that influence drug absorption, distribution, and excretion.[3-5,13]

The New England Journal of Medicine
Downloaded from nejm.org at UNIVERSITA DI TRIESTE on October 28, 2019. For personal use only. No other uses without permission.
Copyright © 2003 Massachusetts Medical Society. All rights reserved.



Polygenic Drug Response

Metabolism genotype		Receptor genotype	Response	
			Efficacy	Toxicity
🟧	+	🟧	65%	Low (5%)
🟧	+	◩	32%	Low
🟧	+	🟦	9%	Low
◩	+	🟧	79%	Moderate (15%)
◩	+	◩	40%	Moderate
◩	+	🟦	10%	Moderate
🟦	+	🟧	80%	High (80%)
🟦	+	◩	40%	High
🟦	+	🟦	10%	High

Figure 1. Polygenic Determinants of Drug Response.

The potential effects of two genetic polymorphisms are illustrated, one involving a drug-metabolizing enzyme (top) and the second involving a drug receptor (middle), depicting differences in drug clearance (or the area under the plasma concentration–time curve [AUC]) and receptor sensitivity in patients who are homozygous for the wild-type allele (WT/WT), are heterozygous for one wild-type and one variant (V) allele (WT/V), or have two variant alleles (V/V) for the two polymorphisms. At the bottom are shown the nine potential combinations of drug-metabolism and drug-receptor genotypes and the corresponding drug-response phenotypes calculated from data at the top, yielding therapeutic indexes (efficacy:toxicity ratios) ranging from 13 (65 percent:5 percent) to 0.125 (10 percent:80 percent).

DRUG METABOLISM

There are more than 30 families of drug-metabolizing enzymes in humans,[3,14] and essentially all have genetic variants, many of which translate into functional changes in the proteins encoded. These monogenic traits are discussed by Weinshilboum.[12] But there is an instructive example of a multigenic effect involving the CYP3A family of P-450 enzymes. About three quarters of whites and half of blacks have a genetic inability to express functional CYP3A5.[15] The lack of functional CYP3A5 may not be readily evident, because many medications metabolized by CYP3A5 are also metabolized by the universally expressed CYP3A4. For medications that are equally metabolized by both enzymes, the net rate of metabolism is the sum of that due to CYP3A4 and that due to CYP3A5; the existence of this dual pathway partially obscures the clinical effects of ge-

The New England Journal of Medicine
Downloaded from nejm.org at UNIVERSITA DI TRIESTE on October 28, 2019. For personal use only. No other uses without permission.
Copyright © 2003 Massachusetts Medical Society. All rights reserved.

netic polymorphism of CYP3A5 but contributes to the large range of total CYP3A activity in humans (Fig. 2). The CYP3A pathway of drug elimination is further confounded by the presence of single-nucleotide polymorphisms in the *CYP3A4* gene that alter the activity of this enzyme for some substrates but not for others.[16] The genetic basis of CYP3A5 deficiency is predominantly a single-nucleotide polymorphism in intron 3 that creates a cryptic splice site causing 131 nucleotides of the intronic sequence to be inserted into the RNA, introducing a termination codon that prematurely truncates the CYP3A5 protein.[15] Although it is now possible to determine which patients express both functional enzymes (i.e., CYP3A4 and CYP3A5), the clinical importance of these variants for the many drugs metabolized by CYP3A remains unclear.

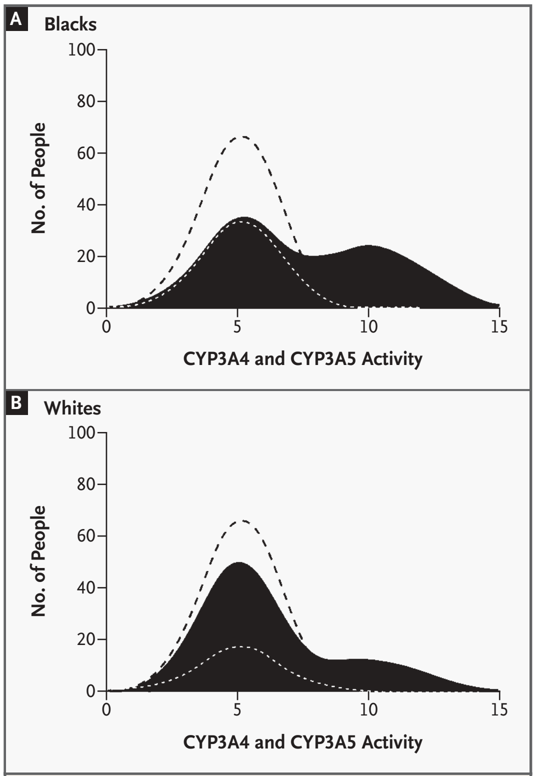


Figure 2. Simulated Activities of Cytochromes P-450 CYP3A4 and CYP3A5 in Blacks and Whites.

The simulated activities of CYP3A4 (black dashed lines) and CYP3A5 (white dashed lines) are shown in blacks (Panel A) and whites (Panel B), assuming a normal distribution and a 10-fold range in activity (shown in arbitrary units) among those expressing functional forms of these enzymes, and further assuming that all patients express CYP3A4, but that only 25 percent of whites and 50 percent of blacks express functional CYP3A5 because of genetic polymorphism. The solid area reflects the combined activity of CYP3A4 and CYP3A5 in the two populations for medications that are metabolized equally by the two enzymes.

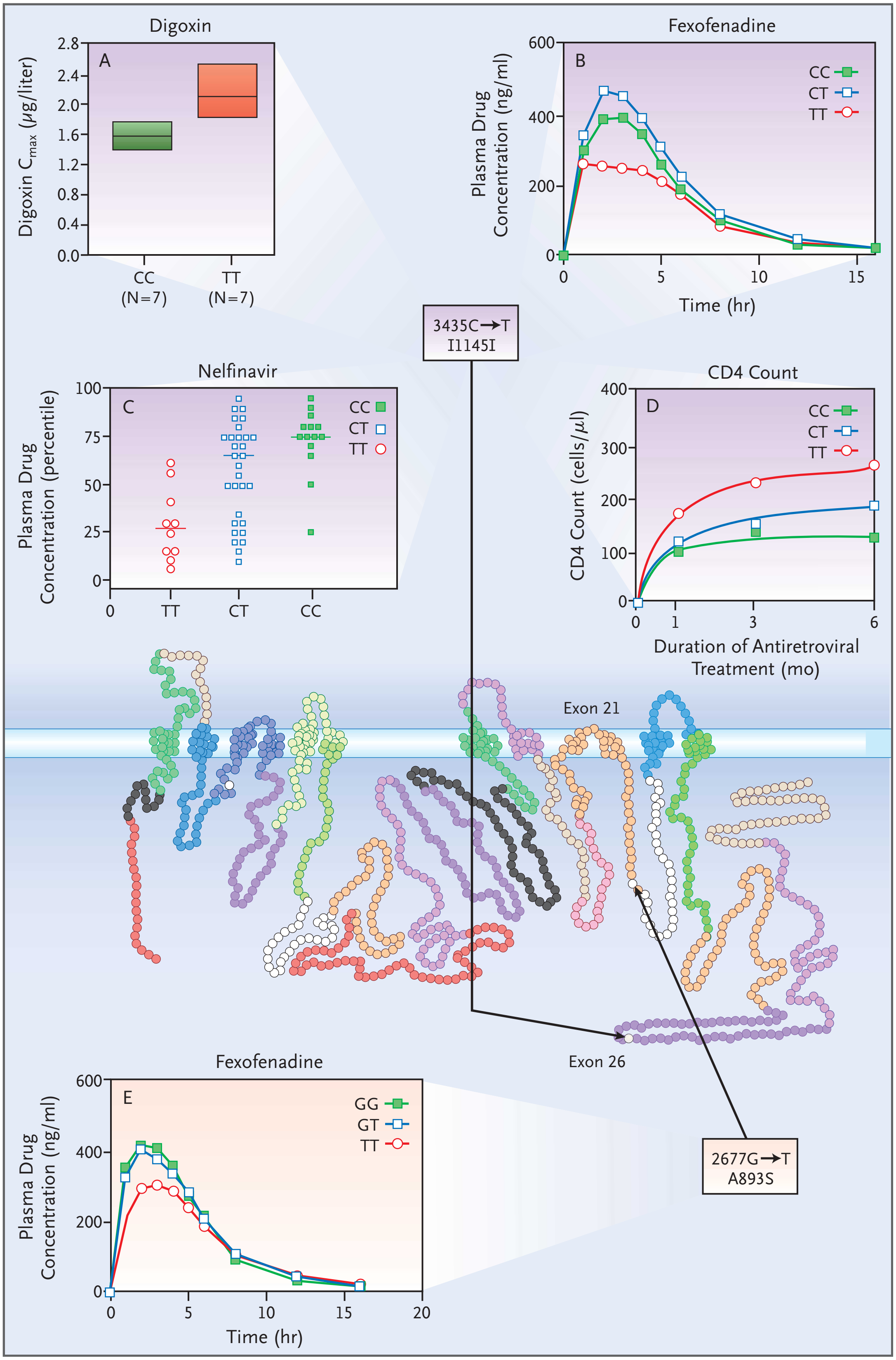
DRUG TRANSPORTERS

Transport proteins have an important role in regulating the absorption, distribution, and excretion of many medications. Members of the adenosine triphosphate (ATP)–binding cassette family of membrane transporters[17] are among the most extensively studied transporters involved in drug disposition and effects. A member of the ATP-binding cassette family, P-glycoprotein, is encoded by the human *ABCB1* gene (also called *MDR1*). A principal function of P-glycoprotein is the energy-dependent cellular efflux of substrates, including bilirubin, several anticancer drugs, cardiac glycosides, immunosuppressive agents, glucocorticoids, human immunodeficiency virus (HIV) type 1 protease inhibitors, and many other medications (Fig. 3).[17,21,22] The expression of P-glycoprotein in many normal tissues suggests that it has a role in the excretion of xenobiotics and metabolites into urine, bile, and the intestinal lumen.[23,24] At the blood–brain barrier, P-glycoprotein in the choroid plexus limits the accumula-

Figure 3 (facing page). Functional Consequences of Genetic Polymorphisms in the Human P-Glycoprotein Transporter Gene *ABCB1* (or *MDR1*).

The schematic diagram of the human P-glycoprotein was adapted from Kim et al.,[18] with each circle representing an amino acid and each color a different exon encoding the corresponding amino acids. Two single-nucleotide polymorphisms in the human *ABCB1* gene have been associated with altered drug disposition (Panels A, B, C, and E) or altered drug effects (Panel D). The synonymous single-nucleotide polymorphism (a single-nucleotide polymorphism that does not alter the amino acid encoded) in exon 26 (the 3435C→T single-nucleotide polymorphism) has been associated with higher oral bioavailability of digoxin in patients homozygous for the T nucleotide (Panel A [C_{max} denotes maximal concentration])[19] but lower plasma concentrations after oral doses of fexofenadine (Panel B)[18] and nelfinavir (Panel C).[20] This single-nucleotide polymorphism has also been linked to better CD4 cell recovery in HIV-infected patients who are treated with nelfinavir and other antiretroviral agents (Panel D).[20] The single-nucleotide polymorphism at nucleotide 2677 (G→T) has been associated with lower plasma fexofenadine concentrations in patients homozygous for the T nucleotide at position 2677 (Panel E).[18] The panels have been adapted from Kim et al.,[18] Hoffmeyer et al.,[19] and Fellay et al.[20]

The New England Journal of Medicine
Downloaded from nejm.org at UNIVERSITA DI TRIESTE on October 28, 2019. For personal use only. No other uses without permission.
Copyright © 2003 Massachusetts Medical Society. All rights reserved.



The New England Journal of Medicine
Downloaded from nejm.org at UNIVERSITA DI TRIESTE on October 28, 2019. For personal use only. No other uses without permission.
Copyright © 2003 Massachusetts Medical Society. All rights reserved.

tion of many drugs in the brain, including digoxin, ivermectin, vinblastine, dexamethasone, cyclosporine, domperidone, and loperamide.[23-25] A synonymous single-nucleotide polymorphism (i.e., a single-nucleotide polymorphism that does not alter the amino acid encoded) in exon 26 (3435C→T) has been associated with variable expression of P-glycoprotein in the duodenum; in patients homozygous for the T allele, duodenal expression of P-glycoprotein was less than half that in patients with the CC genotype.[19] CD56+ natural killer cells from subjects homozygous for 3435C demonstrated significantly lower ex vivo retention of the P-glycoprotein substrate rhodamine (i.e., higher P-glycoprotein function).[26] Digoxin, another P-glycoprotein substrate, has significantly higher bioavailability in subjects with the 3435TT genotype.[19,27] As is typical for many pharmacogenetic traits, there is considerable racial variation in the frequency of the 3435C→T single-nucleotide polymorphism.[28-30]

The 3435C→T single-nucleotide polymorphism is in linkage disequilibrium with a nonsynonymous single-nucleotide polymorphism (i.e., one causing an amino acid change) in exon 21 (2677G→T, leading to Ala893Ser) that alters P-glycoprotein function.[18] Because these two single-nucleotide polymorphisms travel together, it is unclear whether the 3435C→T polymorphism is of functional importance or is simply linked with the causative polymorphism in exon 21. The 2677G→T single-nucleotide polymorphism has been associated with enhanced P-glycoprotein function in vitro and lower plasma fexofenadine concentrations in humans,[18] effects opposite to those reported with digoxin.[27]

The associations between treatment outcome and genetic variants in *CYP3A4*, *CYP3A5*, *CYP2D6*, *CYP2C19*, the chemokine receptor gene *CCR5*, and *ABCB1* have been examined in HIV-infected patients receiving combination antiretroviral therapy with either a protease inhibitor or a nonnucleoside reverse-transcriptase inhibitor.[20] The *ABCB1* 3435C→T polymorphism was associated with significant differences in the plasma pharmacokinetics of nelfinavir (Fig. 3) and efavirenz. Recovery of the CD4 cell count was significantly greater and more rapid in patients with the TT genotype than in patients with either the CT or the CC genotype (Fig. 3). Of many variables evaluated, only the *ABCB1* genotype and the base-line number of HIV RNA copies were significant predictors of CD4 recovery.[20] However, the *ABCB1* 2677G→T single-nucleotide polymorphism was not genotyped, so it remains unclear whether the 3435C→T polymorphism is causative or is simply linked with another polymorphism that is causative.

This example illustrates a common problem in association studies, namely, biologic plausibility. It is not obvious how greater efficacy (CD4 recovery) could be linked to a single-nucleotide polymorphism associated with lower plasma drug concentrations, unless there are specific effects of the *ABCB1* polymorphisms that cause decreased drug efflux from CD4 leukocytes. Overexpression of the gene for another ABC transporter (*ABCC4*, or *MRP4*) confers resistance to some nucleoside antiretroviral agents (e.g., zidovudine).[31] Despite the uncertainty about the mechanisms involved, the clinical value is that a host genetic marker can predict immune recovery after the initiation of antiretroviral treatment and, if validated, may offer a new strategy in tailoring HIV therapy.

GENETIC POLYMORPHISM OF DRUG TARGETS

Genetic variation in drug targets (e.g., receptors) can have a profound effect on drug efficacy, with over 25 examples already identified (Table 1).[3-5] Sequence variants with a direct effect on response occur in the gene for the β_2-adrenoreceptor, affecting the response to β_2-agonists[43,44]; arachidonate 5-lipoxygenase (ALOX5), affecting the response to ALOX5 inhibitors[42]; and angiotensin-converting enzyme (ACE), affecting the renoprotective actions of ACE inhibitors.[32] Genetic differences may also have indirect effects on drug response that are unrelated to drug metabolism or transport, such as methylation of the methylguanine methyltransferase (MGMT) gene promoter, which alters the response of gliomas to treatment with carmustine.[63] The mechanism of this effect is related to a decrease in the efficiency of repair of alkylated DNA in patients with methylated MGMT. It is critical to distinguish this target mechanism from genetic polymorphisms in drug-metabolizing enzymes that affect response by altering drug concentrations, such as the thiopurine methyltransferase polymorphism associated with the hematopoietic toxicity of mercaptopurine[64-66] and susceptibility to radiation-induced brain tumors.[67]

The β_2-adrenoreceptor (coded by the *ADRB2* gene) illustrates another link between genetic polymorphisms in drug targets and clinical responses. Genetic polymorphism of the β_2-adrenoreceptor can alter the process of signal transduction by these receptors.[43,44] Three single-nucleotide polymor-

The New England Journal of Medicine
Downloaded from nejm.org at UNIVERSITA DI TRIESTE on October 28, 2019. For personal use only. No other uses without permission.
Copyright © 2003 Massachusetts Medical Society. All rights reserved.

Table 1. **Genetic Polymorphisms in Drug Target Genes That Can Influence Drug Response.***

Gene or Gene Product	Medication	Drug Effect Associated with Polymorphism
ACE	ACE inhibitors (e.g., enalapril) Fluvastatin	Renoprotective effects, blood-pressure reduction, reduction in left ventricular mass, endothelial function[32-40] Lipid changes (e.g., reductions in low-density lipoprotein cholesterol and apolipoprotein B); progression or regression of coronary atherosclerosis[41]
Arachidonate 5-lipoxygenase	Leukotriene inhibitors	Improvement in FEV_1[42]
β_2-Adrenergic receptor	β_2-Agonists (e.g., albuterol)	Bronchodilatation, susceptibility to agonist-induced desensitization, cardiovascular effects[43-50]
Bradykinin B2 receptor	ACE inhibitors	ACE-inhibitor–induced cough[51]
Dopamine receptors (D2, D3, D4)	Antipsychotics (e.g. haloperidol, clozapine)	Antipsychotic response (D2, D3, D4), antipsychotic-induced tardive dyskinesia (D3), antipsychotic-induced acute akathisia (D3)[52-56]
Estrogen receptor-α	Conjugated estrogens Hormone-replacement therapy	Increase in bone mineral density[57] Increase in high-density lipoprotein cholesterol[58]
Glycoprotein IIIa subunit of glycoprotein IIb/IIIa	Aspirin or glycoprotein IIb/IIIa inhibitors	Antiplatelet effect[59]
Serotonin (5-hydroxytryptamine) transporter	Antidepressants (e.g., clomipramine, fluoxetine, paroxetine)	5-Hydroxytryptamine neurotransmission, antidepressant response[60-62]

* The examples shown are illustrative and not representative of all published studies, which exceed the scope of this review. ACE denotes angiotensin-converting enzyme, and FEV_1 forced expiratory volume in one second.

phisms in *ADRB2* have been associated with altered expression, down-regulation, or coupling of the receptor in response to β_2-adrenoreceptor agonists.[43] Single-nucleotide polymorphisms resulting in an Arg-to-Gly amino acid change at codon 16 and a Gln-to-Glu change at codon 27 are relatively common, with allele frequencies of 0.4 to 0.6, and are under intensive investigation for their clinical relevance.

A recent study of agonist-mediated vasodilatation and desensitization[44] revealed that patients who were homozygous for Arg at *ADRB2* codon 16 had nearly complete desensitization after continuous infusion of isoproterenol, with venodilatation decreasing from 44 percent at base line to 8 percent after 90 minutes of infusion (Fig. 4). In contrast, patients homozygous for Gly at codon 16 had no significant change in venodilatation, regardless of their codon 27 status. Polymorphism at codon 27 was also of functional relevance; subjects homozygous for the Glu allele had higher maximal venodilatation in response to isoproterenol than those with the codon 27 Gln genotype, regardless of their codon 16 status (Fig. 4).[44]

These results are generally consistent with those of studies showing that the forced expiratory volume in one second (FEV_1) after a single oral dose of albuterol was higher by a factor of 6.5 in patients with the Arg/Arg genotype at codon 16 of *ADRB2* than in those with the Gly/Gly genotype (Fig. 4).[48] However, the influence of this genotype was different in patients receiving long-term, regularly scheduled therapy with inhaled β-agonists. Among these patients, those with the Arg/Arg genotype had a gradual decline in the morning peak expiratory flow measured before they had used medication, whereas no change was observed in patients with the Gly/Gly genotype.[47] In addition, the morning peak expiratory flow deteriorated dramatically after the cessation of therapy in patients with the Arg/Arg genotype, but not in those with the Gly/Gly genotype.[47] These data suggest that a codon 16 Arg/Arg genotype may identify patients at risk for deleterious or nonbeneficial effects of regularly scheduled therapy with inhaled β-agonists; the data also suggest that these patients may be candidates for alternative schedules of therapy, earlier initiation of antiinflammatory agents, or both. These findings are also consistent with the aforementioned desensitization of the β_2-adrenoreceptor in patients with a codon 16 Arg/Arg genotype.[44]

At least 13 distinct single-nucleotide polymorphisms have been identified in *ADRB2*.[46] This finding has led to evaluation of the importance of hap-

The New England Journal of Medicine
Downloaded from nejm.org at UNIVERSITA DI TRIESTE on October 28, 2019. For personal use only. No other uses without permission.
Copyright © 2003 Massachusetts Medical Society. All rights reserved.

lotype structure as compared with individual single-nucleotide polymorphisms in determining receptor function and pharmacologic response. Among 77 white, black, Asian, and Hispanic subjects, only 12 distinct haplotypes of the 8192 possible *ADRB2* haplotypes were actually observed.[46] The bronchodilator response to inhaled β-agonist therapy in patients with asthma revealed a stronger association between bronchodilator response and haplotype than between bronchodilator response and any single-nucleotide polymorphism alone.[46] This is not surprising, because haplotype structure is often a better predictor of phenotypic consequences than are individual polymorphisms. This result suggests that it would be desirable to develop simple but robust molecular methods to determine the haplotype structure of patients.[68]

GENETIC POLYMORPHISMS WITH INDIRECT EFFECTS ON DRUG RESPONSE

Polymorphisms in genes encoding proteins that are neither direct targets of medications nor involved in their disposition have been shown to alter the response to treatment in certain situations (Table 2). For example, inherited differences in coagulation factors can predispose women taking oral contraceptives to deep-vein or cerebral-vein thrombosis,[80] whereas polymorphisms in the gene for the cholesterol ester transfer protein have been linked to the progression of atherosclerosis with pravastatin therapy.[75]

Genetic variation in cellular ion transporters can also have an indirect role in predisposing patients to toxic effects of drugs. For example, patients with variant alleles for sodium or potassium transporters may have substantial morbidity or mortality resulting from drug-induced long-QT syndrome. A mutation in *KCNE2*, the gene for an integral membrane subunit that assembles with HERG to form I_{Kr} potassium channels, was identified in a patient who had cardiac arrhythmia after receiving clarithromycin.[76] Additional *KCNE2* variants have been associated with the development of a very long QT interval after therapy with trimethoprim–sulfamethoxazole, with sulfamethoxazole inhibiting potassium channels encoded by the *KCNE2* (8T→A) variant.[77] Because *KCNE2* variants occur in about 1.6 percent of the population and their effect on drug actions can cause death, they are excellent candidates for polygenic strategies to prevent serious drug-induced toxic effects.

Genetic polymorphism in the apolipoprotein E (*APOE*) gene appears to have a role in predicting responses to therapy for Alzheimer's disease and to lipid-lowering drugs.[70,71,82,83] There are numerous allelic variants of the human *APOE* gene (e.g., *APOE* $\epsilon 3$, *APOE* $\epsilon 4$, *APOE* $\epsilon 5$, etc.), which contain one or more single-nucleotide polymorphisms that alter the amino acid sequence of the encoded protein (e.g., apolipoprotein $\epsilon 4$ has a Cys112Arg change). In a study of treatment of Alzheimer's disease with tacrine, 83 percent of the patients without any *APOE* $\epsilon 4$ allele showed improvement in total response and cognitive response after 30 weeks, as compared with 40 percent of patients with at least one *APOE* $\epsilon 4$ allele.[72] However, the greatest individual improvement in this study was seen in a patient with a single *APOE* $\epsilon 4$ allele, the unfavorable genotype, illustrating that a single gene will not always predict the response to a given treatment.[72] Follow-up studies indicate that the interaction between tacrine treatment and *APOE* genotype was strongest for women, again suggesting that many genes are involved in determining the efficacy of a treatment.[84]

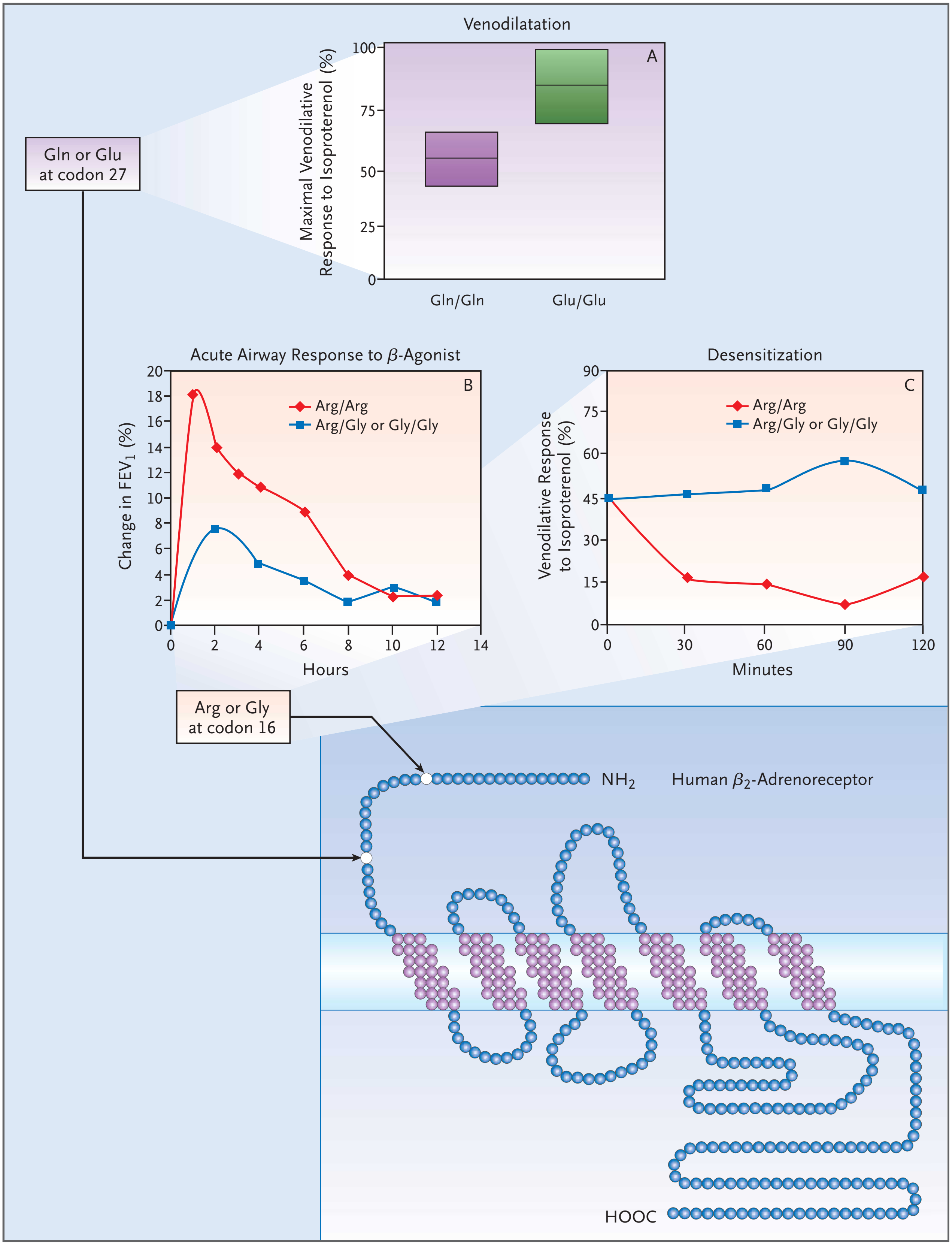
The molecular basis for an association between apolipoprotein genotype and tacrine efficacy has not been elucidated, but it has been postulated that the *APOE* $\epsilon 4$ genotype may have an effect on cholinergic dysfunction in Alzheimer's disease that cannot be consistently overcome by therapy with acetylcholinesterase inhibitors such as tacrine. A randomized, placebo-controlled study of the noradrenergic vasopressinergic agonist S12024 in patients with Alzheimer's disease found the greatest protection of cognition in patients with the *APOE* $\epsilon 4$ genotype.[85] Confirmation of these results may offer an approach to the selection of initial therapy for Alzheimer's disease, with S12024 or similar medications being recommended for patients carrying an *APOE* $\epsilon 4$ allele.

Both phenotypic analysis and genotypic analysis of the *APOE* polymorphism have shown an asso-

Figure 4 (facing page). Functional Consequence of Genetic Polymorphisms in the β_2-Adrenoreceptor (Coded by the *ADRB2* Gene) at Codons 16 and 27.

A homozygous Glu genotype at codon 27 is associated with greater venodilatation after the administration of isoproterenol (Panel A).[44] A homozygous Arg genotype at codon 16 is associated with greater airway response to oral albuterol (Panel B)[48] and greater desensitization to isoproterenol (Panel C).[44] FEV_1 denotes forced expiratory volume in one second.

The New England Journal of Medicine
Downloaded from nejm.org at UNIVERSITA DI TRIESTE on October 28, 2019. For personal use only. No other uses without permission.
Copyright © 2003 Massachusetts Medical Society. All rights reserved.



The New England Journal of Medicine
Downloaded from nejm.org at UNIVERSITA DI TRIESTE on October 28, 2019. For personal use only. No other uses without permission.
Copyright © 2003 Massachusetts Medical Society. All rights reserved.

Table 2. Genetic Polymorphisms in Disease-Modifying or Treatment-Modifying Genes That Can Influence Drug Response.*

Gene or Gene Product	Disease or Response Association	Medication	Influence of Polymorphism on Drug Effect or Toxicity
Adducin	Hypertension	Diuretics	Myocardial infarction or strokes[69]
Apolipoprotein E (APOE)	Progression of atherosclerosis, ischemic cardiovascular events	Statins (e.g., simvastatin)	Enhanced survival[70,71]
Apolipoprotein E (APOE)	Alzheimer's disease	Tacrine	Clinical improvement[72]
HLA	Toxicity	Abacavir	Hypersensitivity reaction[73,74]
Cholesterol ester transfer protein (CETP)	Progression of atherosclerosis	Statins (e.g., pravastatin)	Slowing of progression of atherosclerosis by pravastatin[75]
Ion channels (HERG, KvLQT1, Mink, MiRP1)	Congenital long-QT syndrome	Erythromycin, terfenadine, cisapride, clarithromycin, quinidine	Increased risk of drug-induced torsade de pointes[76-78]
Methylguanine methyltransferase (MGMT)	Glioma	Carmustine	Response of glioma to carmustine[63]
Parkin	Parkinson's disease	Levodopa	Clinical improvement and levodopa-induced dyskinesias[79]
Prothrombin and factor V	Deep-vein thrombosis and cerebral-vein thrombosis	Oral contraceptives	Increased risk of deep-vein and cerebral-vein thrombosis with oral contraceptives[80]
Stromelysin-1	Atherosclerosis progression	Statins (e.g., pravastatin)	Reduction in cardiovascular events by pravastatin (death, myocardial infarction, stroke, angina, and others); reduction in risk of repeated angioplasty[81]

* The examples shown are illustrative and not representative of all published studies, which exceed the scope of this review.

ciation between *APOE* genotype and the response to lipid-lowering medications.[82,86-89] In most studies, patients with an *APOE ϵ2* allele had the greatest diminution of low-density lipoprotein cholesterol after drug therapy. The decrease was greatest for those with *APOE ϵ2*, followed by *APOE ϵ3* and then *APOE ϵ4*. This result was observed after treatment with a diverse range of lipid-lowering agents, including probucol, gemfibrozil, and many different 3-hydroxy-3-methylglutaryl-coenzyme A–reductase inhibitors (statins).[83] However, a significant effect of *APOE* genotype on the response to lipid-lowering agents has not been observed in all studies.[83]

In addition, although the *APOE4* allele was associated with less reduction in total and low-density lipoprotein cholesterol and a smaller increase in high-density lipoprotein cholesterol after fluvastatin therapy, there was no apparent influence of genotype on the progression of coronary artery disease or the incidence of clinical events.[88] Thus, prospective clinical evaluations with robust clinical end points and sufficient sample sizes are needed to define better the usefulness of the *APOE* genotype in selecting the treatment of hyperlipidemia and cardiovascular disease. The potential usefulness of the *APOE* genotype in predicting treatment response must be balanced by the concern that it could be used by insurance companies, health systems, and others to identify those at high risk for Alzheimer's disease, coronary artery disease, and possibly other illnesses.[82]

MOLECULAR DIAGNOSTIC METHODS FOR OPTIMIZING DRUG THERAPY

The potential is enormous for pharmacogenomics to yield a powerful set of molecular diagnostic methods that will become routine tools with which clinicians will select medications and drug doses for individual patients. A patient's genotype needs to be determined only once for any given gene, because except for rare somatic mutations, it does not change. Genotyping methods are improving so rapidly that it will soon be simple to test for thousands of single-nucleotide polymorphisms in one assay. It may be possible to collect a single blood sample from a patient, submit a small aliquot for analysis of a panel of genotypes (e.g., 20,000 single-nucleotide polymorphisms in 5000 genes), and test for those that are important determinants of drug disposition

The New England Journal of Medicine
Downloaded from nejm.org at UNIVERSITA DI TRIESTE on October 28, 2019. For personal use only. No other uses without permission.
Copyright © 2003 Massachusetts Medical Society. All rights reserved.

and effects. In our opinion, genotyping results will be of greatest clinical value if they are reported and interpreted according to the patient's diagnosis and recommended treatment options.

CHALLENGES FOR THE FUTURE

There are a number of critical issues that must be considered as strategies are developed to elucidate the inherited determinants of drug effects. A formidable one is that the inherited component of the response to drugs is often polygenic (Fig. 1). Approaches for elucidating polygenic determinants of drug response include the use of anonymous single-nucleotide polymorphism maps to perform genome-wide searches for polymorphisms associated with drug effects, and candidate-gene strategies based on existing knowledge of a medication's mechanisms of action and pathways of metabolism and disposition. Both these strategies have potential value and limitations, as shown in previous reviews.[5,90,91] However, the candidate-gene strategy has the advantage of focusing resources on a manageable number of genes and polymorphisms that are likely to be important, and it has produced encouraging results in a number of studies.[20,52] The limitations of this approach are the incompleteness of knowledge of a medication's pharmacokinetics and mechanisms of action. Gene-expression profiling[92,93] and proteomic studies[94] are evolving strategies for identifying genes that may influence drug response.

One of the most important challenges in defining pharmacogenetic traits is the need for well-characterized patients who have been uniformly treated and systematically evaluated to make it possible to quantitate drug response objectively. To this end, the norm should be to obtain genomic DNA from all patients enrolled in clinical drug trials, along with appropriate consent to permit pharmacogenetic studies. Because of marked population heterogeneity, a specific genotype may be important in determining the effects of a medication for one population or disease but not for another; therefore, pharmacogenomic relations must be validated for each therapeutic indication and in different racial and ethnic groups. Remaining cognizant of these caveats will help ensure accurate elucidation of genetic determinants of drug response and facilitate the translation of pharmacogenomics into widespread clinical practice.

Supported in part by grants from the National Institutes of Health (R37 CA36401, R01 CA78224, U01 GM61393, U01 GM61394, and U01 GM63340), Cancer Center support grants (CA21765 and CA091842), a Center of Excellence grant from the State of Tennessee, a grant from the Siteman Cancer Center, and a grant from American Lebanese Syrian Associated Charities.

Dr. Evans became a member of the Clinical Genomics Advisory Board of Merck and a member of the Scientific Advisory Board for Signature Genetics and Gentris after this review was written, and he was formerly a member of the Scientific Advisory Board of PPGX. He currently serves as a consultant to Bristol-Myers Squibb. He holds no equity positions in any of these companies. Dr. Evans's laboratory is supported by National Institutes of Health grants. He receives no research support from public or private companies. Dr. McLeod's laboratory is supported by grants from the National Institutes of Health, as well as by research grants from Novartis Pharmaceuticals and Ortho Clinical Diagnostics for projects that do not overlap directly or indirectly with the contents of this article.

REFERENCES

1. Vesell ES. Pharmacogenetic perspectives gained from twin and family studies. Pharmacol Ther 1989;41:535-52.
2. Kalow W, Tang BK, Endrenyi I. Hypothesis: comparisons of inter- and intra-individual variations can substitute for twin studies in drug research. Pharmacogenetics 1998;8:283-9.
3. Evans WE, Relling MV. Pharmacogenomics: translating functional genomics into rational therapeutics. Science 1999; 286:487-91.
4. Evans WE, Johnson JA. Pharmacogenomics: the inherited basis for interindividual differences in drug response. Annu Rev Genomics Hum Genet 2001;2:9-39.
5. McLeod HL, Evans WE. Pharmacogenomics: unlocking the human genome for better drug therapy. Annu Rev Pharmacol Toxicol 2001;41:101-21.
6. Kalow W. Familial incidence of low pseudocholinesterase level. Lancet 1956;2: 576.
7. Carson PE, Flanagan CL, Ickes CE, Alving AS. Enzymatic deficiency in primaquine-sensitive erythrocytes. Science 1956; 124:484-5.
8. Hughes HB, Biehl JP, Jones AP, Schmidt LH. Metabolism of isoniazid in man as related to the occurrence of peripheral neuritis. Am Rev Tuberc 1954;70:266-73.
9. Evans DAP, Manley KA, McKusick VA. Genetic control of isoniazid metabolism in man. Br Med J 1960;2:485-91.
10. Sachidanandam R, Weissman D, Schmidt SC, et al. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 2001;409:928-33.
11. Yates CR, Krynetski EY, Loennechen T, et al. Molecular diagnosis of thiopurine S-methyltransferase deficiency: genetic basis for azathioprine and mercaptopurine intolerance. Ann Intern Med 1997;126:608-14.
12. Weinshilboum R. Inheritance and drug response. N Engl J Med 2003;348:529-37.
13. Meyer UA. Pharmacogenetics and adverse drug reactions. Lancet 2000;356:1667-71.
14. Ingelman-Sundberg M, Oscarson M, McLellan RA. Polymorphic human cytochrome P450 enzymes: an opportunity for individualized drug treatment. Trends Pharmacol Sci 1999;20:342-9.
15. Kuehl P, Zhang J, Lin Y, et al. Sequence diversity in CYP3A promoters and characterization of the genetic basis for polymorphic CYP3A5 expression. Nat Genet 2001;27: 383-91.
16. Sata F, Sapone A, Elizondo G, et al. CYP3A4 allelic variants with amino acid substitutions in exons 7 and 12: evidence for an allelic variant with altered catalytic activity. Clin Pharmacol Ther 2000;67:48-56.
17. Borst P, Evers R, Kool M, Wijnholds J. A family of drug transporters: the multidrug resistance-associated proteins. J Natl Cancer Inst 2000;92:1295-302.
18. Kim RB, Leake BF, Choo EF, et al. Identification of functionally variant MDR1 alleles

The New England Journal of Medicine
Downloaded from nejm.org at UNIVERSITA DI TRIESTE on October 28, 2019. For personal use only. No other uses without permission.
Copyright © 2003 Massachusetts Medical Society. All rights reserved.

among European Americans and African Americans. Clin Pharmacol Ther 2001;70:189-99.
19. Hoffmeyer S, Burk O, von Richter O, et al. Functional polymorphisms of the human multidrug-resistance gene: multiple se-quence variations and correlation of one allele with P-glycoprotein expression and activity in vivo. Proc Natl Acad Sci U S A 2000;97:3473-8.
20. Fellay J, Marzolini C, Meaden ER, et al. Response to antiretroviral treatment in HIV-1-infected individuals with allelic variants of the multidrug resistance transporter 1: a pharmacogenetics study. Lancet 2002;359:30-6.
21. Choo EF, Leake B, Wandel C, et al. Pharmacological inhibition of P-glycoprotein transport enhances the distribution of HIV-1 protease inhibitors into brain and testes. Drug Metab Dispos 2000;28:655-60.
22. Brinkmann U, Roots I, Eichelbaum M. Pharmacogenetics of the human drug-transporter gene MDR1: impact of polymorphisms on pharmacotherapy. Drug Discov Today 2001;6:835-9.
23. Rao VV, Dahlheimer JL, Bardgett ME, et al. Choroid plexus epithelial expression of MDR1 P glycoprotein and multidrug resistance-associated protein contribute to the blood-cerebrospinal-fluid drug-permeability barrier. Proc Natl Acad Sci U S A 1999;96:3900-5.
24. Thiebaut F, Tsuruo T, Hamada H, Gottesman MM, Pastan I, Willingham MC. Cellular localization of the multidrug-resistance gene product P-glycoprotein in normal human tissues. Proc Natl Acad Sci U S A 1987;84:7735-8.
25. Schinkel AH, Wagenaar E, Mol CA, van Deemter L. P-glycoprotein in the blood-brain barrier of mice influences the brain penetration and pharmacological activity of many drugs. J Clin Invest 1996;97:2517-24.
26. Hitzl M, Drescher S, van der Kuip H, et al. The C3435T mutation in the human MDR1 gene is associated with altered efflux of the P-glycoprotein substrate rhodamine 123 from CD56+ natural killer cells. Pharmacogenetics 2001;11:293-8.
27. Sakaeda T, Nakamura T, Horinouchi M, et al. MDR1 genotype-related pharmacokinetics of digoxin after single oral administration in healthy Japanese subjects. Pharm Res 2001;18:1400-4.
28. Ameyaw MM, Regateiro F, Li T, et al. MDR1 pharmacogenetics: frequency of the C3435T mutation in exon 26 is significantly influenced by ethnicity. Pharmacogenetics 2001;11:217-21.
29. McLeod H. Pharmacokinetic differences between ethnic groups. Lancet 2002;359:78.
30. Schaeffeler E, Eichelbaum M, Brinkmann U, et al. Frequency of C3435T polymorphism of MDR1 gene in African people. Lancet 2001;358:383-4.
31. Schuetz JD, Connelly MC, Sun D, et al. MRP4: a previously unidentified factor in resistance to nucleoside-based antiviral drugs. Nat Med 1999;5:1048-51.
32. Jacobsen P, Rossing K, Rossing P, et al. Angiotensin converting enzyme gene polymorphism and ACE inhibition in diabetic nephropathy. Kidney Int 1998;53:1002-6.
33. Kohno M, Yokokawa K, Minami M, et al. Association between angiotensin-converting enzyme gene polymorphisms and regression of left ventricular hypertrophy in patients treated with angiotensin-converting enzyme inhibitors. Am J Med 1999;106:544-9.
34. Ohmichi N, Iwai N, Uchida Y, Shichiri G, Nakamura Y, Kinoshita M. Relationship between the response to the angiotensin converting enzyme inhibitor imidapril and the angiotensin converting enzyme genotype. Am J Hypertens 1997;10:951-5.
35. Okamura A, Ohishi M, Rakugi H, et al. Pharmacogenetic analysis of the effect of angiotensin-converting enzyme inhibitor on restenosis after percutaneous transluminal coronary angioplasty. Angiology 1999;50:811-22.
36. Penno G, Chaturvedi N, Talmud PJ, et al. Effect of angiotensin-converting enzyme (ACE) gene polymorphism on progression of renal disease and the influence of ACE inhibition in IDDM patients: findings from the EUCLID Randomized Controlled Trial: EURODIAB Controlled Trial of Lisinopril in IDDM. Diabetes 1998;47:1507-11.
37. Perna A, Ruggenenti P, Testa A, et al. ACE genotype and ACE inhibitors induced renoprotection in chronic proteinuric nephropathies. Kidney Int 2000;57:274-81.
38. Prasad A, Narayanan S, Husain S, et al. Insertion-deletion polymorphism of the ACE gene modulates reversibility of endothelial dysfunction with ACE inhibition. Circulation 2000;102:35-41.
39. Sasaki M, Oki T, Iuchi A, et al. Relationship between the angiotensin converting enzyme gene polymorphism and the effects of enalapril on left ventricular hypertrophy and impaired diastolic filling in essential hypertension: M-mode and pulsed Doppler echocardiographic studies. J Hypertens 1996;14:1403-8.
40. Stavroulakis GA, Makris TK, Krespi PG, et al. Predicting response to chronic antihypertensive treatment with fosinopril: the role of angiotensin-converting enzyme gene polymorphism. Cardiovasc Drugs Ther 2000;14:427-32.
41. Marian AJ, Safavi F, Ferlic L, Dunn JK, Gotto AM, Ballantyne CM. Interactions between angiotensin-I converting enzyme insertion/deletion polymorphism and response of plasma lipids and coronary atherosclerosis to treatment with fluvastatin: the Lipoprotein and Coronary Atherosclerosis Study. J Am Coll Cardiol 2000;35:89-95.
42. Drazen JM, Yandava CN, Dube L, et al. Pharmacogenetic association between ALOX5 promoter genotype and the re-sponse to anti-asthma treatment. Nat Genet 1999;22:168-70.
43. Liggett SB. Beta(2)-adrenergic receptor pharmacogenetics. Am J Respir Crit Care Med 2000;161:S197-S201.
44. Dishy V, Sofowora GG, Xie H-G, et al. The effect of common polymorphisms of the β_2-adrenergic receptor on agonist-mediated vascular desensitization. N Engl J Med 2001;345:1030-5.
45. Cockcroft JR, Gazis AG, Cross DJ, et al. Beta(2)-adrenoceptor polymorphism determines vascular reactivity in humans. Hypertension 2000;36:371-5.
46. Drysdale CM, McGraw DW, Stack CB, et al. Complex promoter and coding region beta 2-adrenergic receptor haplotypes alter receptor expression and predict in vivo responsiveness. Proc Natl Acad Sci U S A 2000;97:10483-8.
47. Israel E, Drazen JM, Liggett SB, et al. Effect of polymorphism of the beta(2)-adrenergic receptor on response to regular use of albuterol in asthma. Int Arch Allergy Immunol 2001;124:183-6.
48. Lima JJ, Thomason DB, Mohamed MH, Eberle LV, Self TH, Johnson JA. Impact of genetic polymorphisms of the beta2-adrenergic receptor on albuterol bronchodilator pharmacodynamics. Clin Pharmacol Ther 1999;65:519-25.
49. Martinez FD, Graves PE, Baldini M, Solomon S, Erickson R. Association between genetic polymorphisms of the beta2-adrenoceptor and response to albuterol in children with and without a history of wheezing. J Clin Invest 1997;100:3184-8.
50. Tan S, Hall IP, Dewar J, Dow E, Lipworth B. Association between beta 2-adrenoceptor polymorphism and susceptibility to bronchodilator desensitization in moderately severe stable asthmatics. Lancet 1997;350:995-9.
51. Mukae S, Aoki S, Itoh S, Iwata T, Ueda H, Katagiri T. Bradykinin B(2) receptor gene polymorphism is associated with angiotensin-converting enzyme inhibitor-related cough. Hypertension 2000;36:127-31.
52. Arranz MJ, Munro J, Birkett J, et al. Pharmacogenetic prediction of clozapine response. Lancet 2000;355:1615-6.
53. Basile VS, Masellis M, Badri F, et al. Association of the MscI polymorphism of the dopamine D3 receptor gene with tardive dyskinesia in schizophrenia. Neuropsychopharmacology 1999;21:17-27.
54. Eichhammer P, Albus M, Borrmann-Hassenbach M, et al. Association of dopamine D3-receptor gene variants with neuroleptic induced akathisia in schizophrenic patients: a generalization of Steen's study on DRD3 and tardive dyskinesia. Am J Med Genet 2000;96:187-91.
55. Hwu HG, Hong CJ, Lee YL, Lee PC, Lee SF. Dopamine D4 receptor gene polymorphisms and neuroleptic response in schizophrenia. Biol Psychiatry 1998;44:483-7.
56. Kaiser R, Konneker M, Henneken M, et al. Dopamine D4 receptor 48-bp repeat polymorphism: no association with response to antipsychotic treatment, but association with catatonic schizophrenia. Mol Psychiatry 2000;5:418-24.
57. Ongphiphadhanakul B, Chanprasertyothin S, Payatikul P, et al. Oestrogen-

The New England Journal of Medicine
Downloaded from nejm.org at UNIVERSITA DI TRIESTE on October 28, 2019. For personal use only. No other uses without permission.
Copyright © 2003 Massachusetts Medical Society. All rights reserved.

receptor-alpha gene polymorphism affects response in bone mineral density to oestrogen in post-menopausal women. Clin Endocrinol (Oxf) 2000;52:581-5.

58. Herrington DM, Howard TD, Hawkins GA, et al. Estrogen-receptor polymorphisms and effects of estrogen replacement on high-density lipoprotein cholesterol in women with coronary disease. N Engl J Med 2002; 346:967-74.

59. Michelson AD, Furman MI, Goldschmidt-Clermont P, et al. Platelet GP IIIa PI(A) polymorphisms display different sensitivities to agonists. Circulation 2000;101: 1013-8.

60. Kim DK, Lim SW, Lee S, et al. Serotonin transporter gene polymorphism and antidepressant response. Neuroreport 2000;11: 215-9.

61. Smeraldi E, Zanardi R, Benedetti F, Di Bella D, Perez J, Catalano M. Polymorphism within the promoter of the serotonin transporter gene and antidepressant efficacy of fluvoxamine. Mol Psychiatry 1998;3:508-11.

62. Whale R, Quested DJ, Laver D, Harrison PJ, Cowen PJ. Serotonin transporter (5-HTT) promoter genotype may influence the prolactin response to clomipramine. Psychopharmacology (Berl) 2000;150:120-2.

63. Esteller M, Garcia-Foncillas J, Andion E, et al. Inactivation of the DNA-repair gene *MGMT* and the clinical response of gliomas to alkylating agents. N Engl J Med 2000;343: 1350-4. [Erratum, N Engl J Med 2000;343: 1740.]

64. Evans WE, Hon YY, Bomgaars L, et al. Preponderance of thiopurine S-methyltransferase deficiency and heterozygosity among patients intolerant to mercaptopurine or azathioprine. J Clin Oncol 2001;19:2293-301.

65. Black AJ, McLeod HL, Capell HA, et al. Thiopurine methyltransferase genotype predicts therapy-limiting severe toxicity from azathioprine. Ann Intern Med 1998; 129:716-8.

66. Relling MV, Hancock ML, Rivera GK, et al. Mercaptopurine therapy intolerance and heterozygosity at the thiopurine S-methyltransferase gene locus. J Natl Cancer Inst 1999;91:2001-8.

67. Relling MV, Rubnitz JE, Rivera GK, et al. High incidence of secondary brain tumours after radiotherapy and antimetabolites. Lancet 1999;354:34-9.

68. McDonald OG, Krynetski EY, Evans WE. Molecular haplotyping of genomic DNA for multiple single-nucleotide polymorphisms located kilobases apart using long-range polymerase chain reaction and intramolecular ligation. Pharmacogenetics 2002;12:93-9.

69. Psaty BM, Smith NL, Heckbert SR, et al. Diuretic therapy, the α-adducin gene variation, and the risk of myocardial infarction or stroke in persons with treated hypertension. JAMA 2002;287:1680-9.

70. Gerdes LU, Gerdes C, Kervinen K, et al. The apolipoprotein epsilon4 allele determines prognosis and the effect on prognosis of simvastatin in survivors of myocardial infarction: a substudy of the Scandinavian Simvastatin Survival Study. Circulation 2000;101:1366-71.

71. Ordovas JM, Lopez-Miranda J, Perez-Jimenez F, et al. Effect of apolipoprotein E and A-IV phenotypes on the low density lipoprotein response to HMG CoA reductase inhibitor therapy. Atherosclerosis 1995; 113:157-66.

72. Poirier J, Delisle MC, Quirion R, et al. Apolipoprotein E4 allele as a predictor of cholinergic deficits and treatment outcome in Alzheimer disease. Proc Natl Acad Sci U S A 1995;92:12260-4.

73. Mallal S, Nolan D, Witt C, et al. Association between presence of HLA-B*5701, HLA-DR7, and HLA-DQ3 and hypersensitivity to HIV-1 reverse-transcriptase inhibitor abacavir. Lancet 2002;359:727-32.

74. Hetherington S, Hughes AR, Mosteller M, et al. Genetic variations in HLA-B region and hypersensitivity reaction to abacavir. Lancet 2002;359:1121-2.

75. Kuivenhoven JA, Jukema JW, Zwinderman AH, et al. The role of a common variant of the cholesteryl ester transfer protein gene in the progression of coronary atherosclerosis. N Engl J Med 1998;338:86-93.

76. Abbott GW, Sesti F, Splawski I, et al. MiRP1 forms IKr potassium channels with HERG and is associated with cardiac arrhythmia. Cell 1999;97:175-87.

77. Sesti F, Abbott GW, Wei J, et al. A common polymorphism associated with antibiotic-induced cardiac arrhythmia. Proc Natl Acad Sci U S A 2000;97:10613-8.

78. Napolitano C, Schwartz PJ, Brown AM, et al. Evidence for a cardiac ion channel mutation underlying drug-induced QT prolongation and life-threatening arrhythmias. J Cardiovasc Electrophysiol 2000;11:691-6.

79. Lücking CB, Dürr A, Bonifati V, et al. Association between early-onset Parkinson's disease and mutations in the *parkin* gene. N Engl J Med 2000;342:1560-7.

80. Martinelli I, Sacchi E, Landi G, Taioli E, Duca F, Mannucci PM. High risk of cerebral-vein thrombosis in carriers of a prothrombin-gene mutation and in users of oral contraceptives. N Engl J Med 1998;338:1793-7.

81. de Maat MP, Jukema JW, Ye S, et al. Effect of the stromelysin-1 promoter on efficacy of pravastatin in coronary atherosclerosis and restenosis. Am J Cardiol 1999;83: 852-6.

82. Issa AM, Keyserlingk EW. Apolipoprotein E genotyping for pharmacogenetic purposes in Alzheimer's disease: emerging ethical issues. Can J Psychiatry 2000;45:917-22.

83. Siest G, Bertrand P, Herbeth B, et al. Apolipoprotein E polymorphisms and concentration in chronic diseases and drug responses. Clin Chem Lab Med 2000;38: 841-52.

84. Farlow MR, Lahiri DK, Poirier J, Davignon J, Schneider L, Hui SL. Treatment outcome of tacrine therapy depends on apolipoprotein genotype and gender of the subjects with Alzheimer's disease. Neurology 1998;50:669-77.

85. Richard F, Helbecque N, Neuman E, Guez D, Levy R, Amouyel P. APOE genotyping and response to drug treatment in Alzheimer's disease. Lancet 1997;349:539.

86. Nestruck AC, Bouthillier D, Sing CF, Davignon J. Apolipoprotein E polymorphism and plasma cholesterol response to probucol. Metabolism 1987;36:743-7.

87. Pedro-Botet J, Schaefer EJ, Bakker-Arkema RG, et al. Apolipoprotein E genotype affects plasma lipid response to atorvastatin in a gender specific manner. Atherosclerosis 2001;158:183-93.

88. Watanabe J, Kobayashi K, Umeda F, et al. Apolipoprotein E polymorphism affects the response to pravastatin on plasma apolipoproteins in diabetic patients. Diabetes Res Clin Pract 1993;20:21-7.

89. Ballantyne CM, Herd JA, Stein EA, et al. Apolipoprotein E genotypes and response of plasma lipids and progression-regression of coronary atherosclerosis to lipid-lowering drug therapy. J Am Coll Cardiol 2000;36:1572-8.

90. Cargill M, Daley GQ. Mining for SNPs: putting the common variants–common disease hypothesis to the test. Pharmacogenomics 2000;1:27-37.

91. Sham P. Shifting paradigms in gene-mapping methodology for complex traits. Pharmacogenomics 2001;2:195-202.

92. Staunton JE, Slonim DK, Coller HA, et al. Chemosensitivity prediction by transcriptional profiling. Proc Natl Acad Sci U S A 2001;98:10787-92.

93. Yeoh EJ, Ross ME, Shurtleff SA, et al. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. Cancer Cell 2002;1:133-43.

94. Liotta LA, Kohn EC, Petricoin EF. Clinical proteomics: personalized molecular medicine. JAMA 2001;286:2211-4.

Copyright © 2003 Massachusetts Medical Society.

The New England Journal of Medicine
Downloaded from nejm.org at UNIVERSITA DI TRIESTE on October 28, 2019. For personal use only. No other uses without permission.
Copyright © 2003 Massachusetts Medical Society. All rights reserved.

View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Springer - Publisher Connector

Lin and Farrer *Genome Medicine* 2014, **6**:48
http://genomemedicine.com/content/6/6/48

Genome Medicine

REVIEW

Genetics and genomics of Parkinson's disease

Michelle K Lin and Matthew J Farrer*

Abstract

Parkinson's disease (PD) is a progressively debilitating neurodegenerative syndrome. Although best described as a movement disorder, the condition has prominent autonomic, cognitive, psychiatric, sensory and sleep components. Striatal dopaminergic innervation and nigral neurons are progressively lost, with associated Lewy pathology readily apparent on autopsy. Nevertheless, knowledge of the molecular events leading to this pathophysiology is limited. Current therapies offer symptomatic benefit but they fail to slow progression and patients continue to deteriorate. Recent discoveries in sporadic, Mendelian and more complex forms of parkinsonism provide novel insight into disease etiology; 28 genes, including those encoding alpha-synuclein (*SNCA*), leucine-rich repeat kinase 2 (*LRRK2*) and microtubule-associated protein tau (*MAPT*), have been linked and/or associated with PD. A consensus regarding the affected biological pathways and molecular processes has also started to emerge. In early-onset and more a typical PD, deficits in mitophagy pathways and lysosomal function appear to be prominent. By contrast, in more typical late-onset PD, chronic, albeit subtle, dysfunction in synaptic transmission, early endosomal trafficking and receptor recycling, as well as chaperone-mediated autophagy, provide a unifying synthesis of the molecular pathways involved. Disease-modification (neuroprotection) is no longer such an elusive goal given the unparalleled opportunity for diagnosis, translational neuroscience and therapeutic development provided by genetic discovery.

Clinical features, main treatments and challenges

Parkinson's disease (PD) is a progressively debilitating neurodegenerative disease that becomes increasingly challenging to manage. Until recently, the underlying molecular cause(s) have remained elusive, as have hopes of disease-modification (neuroprotection). The need for advances in understanding and treating the disease is great as PD is the second most common neurodegenerative disorder after Alzheimer's disease. The estimated worldwide prevalence is 1% in the population >60 years of age, increasing to 4% at 80 years of age [1]. The median age of onset is about 70 years but about 4% of patients manifest early-onset disease (before they reach 50 years of age).

In PD, the clinical heterogeneity, disease course and response to medication vary widely [2]. Motor symptoms are associated with profound neuronal loss in the *substantia nigra pars compacta* (SN), depleting the striatum of dopaminergic inhibition. Hence, restorative dopaminergic therapies are the main treatment. Monoamine oxidase inhibitors are initially used to prevent endogenous dopamine catabolism; alternatively, L-DOPA, the metabolic precursor of dopamine, and/or dopamine agonists are used. In selected patients, deep brain stimulation (DBS) of striatal output pathways has also proven effective [3]. There are, however, several non-motor symptoms, of which many are non-dopaminergic and without remedy [4]. For example, 30% of patients develop mild-cognitive impairment within 5 years of motor symptoms and many develop dementia [5]. PD is clinically and/or pathologically distinct from other forms of parkinsonism (Box 1). A definitive diagnosis of PD requires the presence of Lewy bodies and Lewy neurites (proteinaceous intracellular inclusions) in the brain stem (midbrain), although these lesions are often more widespread [6,7]. Here, we review recent genetic and genomic findings in studies of PD, we provide some integration and synthesis of the molecular pathways involved, and we discuss the translational implications.

Early linkage and candidate gene studies

Classical linkage analysis has proven a powerful approach for the identification of specific disease-associated genes and mutations in families with multi-incident parkinsonism [8]. Genome-wide association and twin studies further demonstrate that even idiopathic sporadic

---

* Correspondence: mfarrer@can.ubc.ca
Djavad Mowafaghian Centre for Brain Health, Centre for Applied Neurogenetics, Department of Medical Genetics, University of British Columbia, Vancouver, BC V6T 1Z3, Canada

BioMed Central

© 2014 Lin and Farrer; licensee BioMed Central Ltd. The licensee has exclusive rights to distribute this article, in any medium, for 12 months following its publication. After this time, the article is available under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

families with dominant, recessive or X-linked patterns of segregation. Not surprisingly, *in vivo* modeling in mammalian systems is challenging if the expectation is to recapitulate the human phenotype. Gene discovery efforts in PD have been expertly reviewed and we provide some historical context in Box 2. In this review, we focus on more typical, late-onset PD with Lewy body pathology, the disease type suffered by the majority of patients, and suggest how recent discoveries might unify existing ideas to suggest novel pathways and therapeutic targets.

In this context of genomic discovery, genes at three loci - alpha-synuclein (*SNCA*), leucine-rich repeat kinase 2 (*LRRK2*) and microtubule-associated protein tau (*MAPT*) - deserve special mention, although the molecular relationship between them has yet to be elucidated. Both *SNCA* and *LRRK2* assignments were originally implicated by the discovery (by linkage analysis) of pathogenic mutations that segregate within families, and these observations were extended into idiopathic PD by candidate gene studies. Combined pooled analysis by the Genetic Epidemiology of Parkinson's disease Consortium (www.geopd.org), literature meta-analysis (www.pdgene.org) and more recent genome-wide association studies (GWAS) have provided compelling support for the involvement of these loci.

Missense and multiplication mutations (duplication and triplication) in *SNCA* lead to PD, subsequent dementia and fulminant diffuse Lewy body disease on autopsy [11]. Levels of gene expression are inversely correlated with age at symptom onset. In rodents, *SNCA* overexpression may recapitulate many of the features of PD, whereas knockout mice are viable and fertile and appear to have little sign of disease [12]. The alpha-synuclein protein promotes presynaptic SNARE complex assembly, synaptic vesicle exocytosis and reciprocal plasmalemma endocytosis [13]. Alpha-synuclein protein aggregates, however, may also behave pathologically as prion proteins [14]. Transplants of fetal tissue into the striatum of human patients have been observed to develop Lewy-body pathology [15]. Similarly, iatrogenic inoculation of alpha-synuclein oligomers into mouse brain leads to widespread Lewy-like pathology, albeit requiring endogenous alpha-synuclein for transmission [16]. Hence, down-regulation or suppression of *SNCA* may represent one mechanism to slow disease progression [17]. Similarly, therapies that enhance the clearance of Lewy aggregates, including immunotherapies targeting potentially toxic forms of alpha-synuclein, might be neuroprotective [18].

LRRK2 parkinsonism is clinically indistinguishable from idiopathic PD. The age of onset and age-dependent cumulative incidence is similarly broad, although disease progression in carriers of *LRRK2* mutations is more homogeneous [19]. A founder haplotype has been noted for *LRRK2* p.G2019S in most populations. Penetrance appears to be ethnic-specific with a lifetime cumulative

Box 1. Parkinsonism and Parkinson's disease

Parkinsonism is used as an adjective to describe a movement disorder generally consisting of one of more features of tremor, bradykinesia and rigidity. The most common forms are drug-induced (for example, as a side-effect of antipsychotic neuroleptics), or vascular resulting from a stroke. When used as an adjective, the term parkinsonism does not imply that the symptoms are progressive or neurodegenerative, or that they are associated with a specific neuropathology. Parkinsonism may not respond to levodopa therapy. PD itself has the same core triad of symptoms, but it begins asymmetrically, one side of the body being more profoundly affected than the other, and postural instability is typically a feature. In PD, symptoms respond well to levodopa replacement therapy, and typically worsen with disease duration despite medication. Symptoms become bilateral in advanced PD, and dopaminergic imaging, using modalities such as DaTscan, highlights the same pattern of progression.

PD, but not parkinsonism, is often associated with a variety of autonomic problems (constipation, bowel and bladder incontinence, drooling), cognitive dysfunctions (from mild impairments in executive function to dementia), motor problems (from dystonia to dyskinesia), neuropsychiatric symptoms (mood disorders, such as depression to impulse control), sensory problems (from unexplained pain syndromes to hyposmia or anosmia) and sleep disorders (daytime sleepiness to REM sleep behavior disorder). Several of these problems, including dyskinesia, mood disorders and daytime sleepiness, may be associated with levodopa therapy or the use of dopamine agonists.

A definite diagnosis of PD is reserved for patients with Lewy body pathology and neuronal loss in the midbrain, which is often more widely distributed to involve the myenteric plexus, vagus and olfactory bulb. By contrast, parkinsonism may be associated with a variety of pathologies, including neurofibrillary tangles (such as in progressive supranuclear palsy, corticobasal ganglionic degeneration, or parkinsonism-dementia complex of Guam), to the predominant oligodendroglial alpha-synuclein pathology of multiple system atrophy, or predominant cortical Lewy pathology associated with dementia with Lewy bodies.

PD has a significant genetic component [9,10]. Mutant gene discovery by linkage with association provides an unequivocal burden of proof, and is the foundation required for translational neuroscience. Nevertheless, genetically defined carriers may have variable expressivity and penetrance and may never 'phenoconvert' to symptomatic disease. Age remains an important determinant, even in

Box 2. Genetic insights and evolving neuroscience

1997 A missense substitution, p.A53T, is discovered in the gene encoding alpha-synuclein (*SNCA*) in a family from Contursi, Italy whose members were susceptible to autosomal dominant, late-onset parkinsonism [56]. Alpha-synuclein, then known as 'Non-amyloid component of plaques' (NACP), provides a link between PD and Alzheimer's disease. Whether alpha-synuclein monomers, oligomers or fibrils are the toxic species becomes a topic for debate. Physiologically, the Zebra Finch homolog of *SNCA* is found to be required for song learning, pre-synaptic plasticity, and vesicle trafficking in neurotransmission.

Alpha-synuclein immunohistochemistry reveals much more occult Lewy pathology than had been visualized previously by hematoxylin and eosin (H&E) staining, and replaced the use of PGP9.5 (an antibody for Ubiquitin C-terminal hydrolase1 (UCHL1)) [136]. It is argued that familial and idiopathic forms of parkinsonism are the same disease, with similar ontology. Nevertheless, epidemiologists claim that the etiology of late-onset PD is environmental, rather than due to a genetic predisposition, supported by the results of twin studies. Further debate centers on whether ubiqinated Lewy body 'aggresomes' are pathologic or protective.

1998 Homozygous parkin deletions are linked to recessively inherited juvenile and early-onset parkinsonism, albeit without documented Lewy pathology [109]. Parkin mutations soon explain around 15% of all early-onset parkinsonism (in those younger than 45 years) [137]. As the parkin mutations affect a ubiquitin E3 ligase, the field focuses on proteasome inhibition, attempting to identify parkin's substrates and to nominate the toxic species. Patients with parkin mutations and Lewy or tau pathology have subsequently been described [138,139]. The crystal structure of the ubiquitin ligase reveals how its activity is regulated [116].

1999-2001 Candidate gene studies of polymorphic variability in *SNCA* and microtubule-associated protein tau (*MAPT*) highlight the roles of these genes in idiopathic, late-onset PD [57,140,141]. *MAPT* is implicated in progressive supranuclear palsy, in part because of the discovery of splicing mutations in another 4R-tauopathy, frontotemporal dementia with parkinsonism linked to chromosome 17 (FTDP-17) [44]. Large-scale GWAS provide further confirmation [9,52,55,142].

2003 *DJ-1* mutations that were discovered in early-onset parkinsonism [123] highlight a role for oxidative stress in PD. The findings help to justify the use of toxin-based models using 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine (MPTP) as an analog of the herbicide paraquat or the pesticide rotenone.

2003-2004 *SNCA* triplication and duplication mutations demonstrate a dose-dependent relationship between expression and pathogenicity [58,143-146] that is now supported by several mouse models. Lowering *SNCA* RNA and/or SNCA protein expression is nominated as a therapeutic target, supported by *in vivo* studies.

2004 Recessively inherited mutations are identified in the Pten-induced kinase 1 gene (*PINK1*) in early-onset parkinsonism [147]. By 2006, PINK1 was found to regulate Parkin recruitment and mitophagy [114,115]. *FBXO7* is within the same pathway, pointing to mitochondrial maintenance as a therapeutic opportunity [68]. Hexokinase is identified through interaction screens as an upstream component of the pathway [119]. PINK1 is shown to phosphorylate ubiquitin to activate parkin [117].

2004-2005 Dominantly inherited mutations are identified in the *LRRK2* gene in late-onset PD [32]. In 2005, *LRRK2* p.G2019S was linked to PD in Norwegian families [148], and found to be the major determinant of sporadic PD in Ashkenazi Jews and North-African Berbers. Kinase inhibition is nominated as a therapeutic strategy. Pleomorphic alpha-synuclein, 4R-tau or ubiquitin pathologies in affected carriers suggest that Lewy pathology should not be required for a definite diagnosis of PD [32]. LRRK2 is implicated in protein sorting or trafficking and in autophagy. Polymorphic variants in *LRRK2* are found to lower or increase the risk of sporadic PD [34].

2006 *ATP13A2* recessive mutations are identified in juvenile and early-onset parkinsonism [102]. Several lysosome-associated proteins have subsequently been linked to rare and rather atypical forms, including *ATP1A3* in rapid-onset parkinsonism-dystonia, and most recently *ATP6AP2* [71]. Polymorphic variants in the glucocerebrosidase gene *GBA1* are reproducibly associated with late-onset PD, highlighting a role for endosomal trafficking and lysosomal function [98].

2009 GWAS studies in PD confirm associations with *SNCA* and *MAPT*, and find evidence for additional loci [50-52].

2011 *VPS35* D620N is linked to dominant late-onset PD [63]. Many families have the same, albeit *de novo*, substitution that supports pathogenicity. By contrast, mutations in *EIF4G1* R1205H require further genetic and functional support for their assignment [149].

2013-2014 Mutations in *DNAJC13*, along with *VPS35*, in late-onset Lewy-body PD further highlight involvement of the retromer-WASH complex, endosomal protein sorting and trafficking [63,64,93].

2014 Further meta-analysis of GWAS supports known and novel loci. Nevertheless, for most loci the precise gene and underlying variability remain elusive.

risk of parkinsonism in p.G2019S carriers being 22% for Ashkenazi Jews (living in the US), 45% for Norwegians and 80% for Arab-Berbers [20-22], which is an important consideration for genetic counseling. For Ashkenazi Jews in Israel and Arab-Berbers in Tunisia the penetrance of disease is similar. In these countries, the frequency *of LRRK2* p.G2019S carriers in healthy controls is relatively high at 1 to 2% [23,24]. While genetic drift may be sufficient explanation for the high frequency of *LRRK2* p.G2019S carriers in healthy controls, positive selection may also contribute; *LRRK2* is associated with intestinal inflammatory disorders, immune response and kidney function, which if compromised may be most deleterious in hot climates [25-27]. Overall, only a small proportion (approximately 1%) of familial and sporadic PD is likely explained by *LRRK2* p.G2019S; other variability, such as the p.G2385R mutation, which is frequent throughout Asia, most significantly contributes to population-attributable risk [28] (Table 1).

At post-mortem, the majority of patients with pathogenic *LRRK2* mutations have Lewy-body disease [29]. However, even within families with the same mutation, pleomorphic pathologies have been observed. These include neurofibrillary tangles and tufted astrocytes (4R-tauopathies, as inclusion of *MAPT* exon 10 leads to the production of tau protein with four microtubule-binding domains), Tar DNA-binding protein 43 and ubiquitin immune-positive aggregates, or simply nigral neuronal loss with gliosis [30-32]. The variable penetrance and alternative end-stage pathologies most likely reflect genetic and/or environmental modifiers and stochastic factors, and have yet to be defined. Nevertheless, genetically defined cohorts of patients with parkinsonism such as LRRK2 p.G2019S parkinsonism might allow the identification of biomarkers of disease progression and inform clinical trials [19].

As the most common genetic cause of PD, LRRK2 and its protein interactions are a logical place to search for novel therapeutic targets. The domains of LRRK2 include armadillo and ankyrin repeat regions, leucine-rich repeat (LRR), Ras of complex GTPase (Roc), C-terminal of Ras (COR), kinase and WD40. LRRK2 is a ROCO protein, with a Ras GTPase and a kinase in one molecule; these activities have established roles in other organisms or cell types in dynamically modifying the actin cytoskeleton [33]. Pathogenic *LRRK2* mutations are primarily found in the GTPase Roc domain (p.R1437H, p.R1441C/G/H), the kinase domain (p.G2019S, p.I2020T) and intervening C-terminal of Roc (p.Y1699C), whereas susceptibility variants may also be found in protein-protein interaction domains (WD40 p.G2385R) [34]. Competitive inhibition of LRRK2 kinase is presently considered as one therapeutic target given that the p.G2019S mutation activates kinase activity, autophosphorylation and/or phosphorylation of substrates [35]. However, data on the first identified substrate, moesin, which is a filamentous actin tether, have not been recapitulated *ex* or *in vivo* [36]. LRRK2 protein levels also diminish with aging, with knock-in of *LRRK2* pathogenic mutations into the mouse genome [26] and with kinase inhibition [37,38]. LRRK2 probably functions as a dimer or higher molecular weight scaffold, with many protein-protein interactions. The activities of the Roc, COR and kinase domains are interconnected [39], and the many physiologic functions of the protein complex have yet to be fully elucidated.

Table 1 Genomic loci implicated in Parkinson's disease by genome-wide association analyses

Gene	Chr	AssociatedSNP/locus	Genes within locus[a]	Odds ratio[95% CI]	*P*-value
GBA	1q21	N370S	*TRIM46*, *MUC1*, *MIR92B*, *THBS3*, *GBAP1-GBA-FAM189B*, *SCAMP3*, *CLK2*, *HCN3*, *PKLR*	3.37[2.67-4.29]	1.11E-24
SYT11/RAB25	1q21	chr1:154105678	*MIR7851*, *UBQLN4*, *LAMTOR2-RAB25-MEX3A*, *LMNA*	1.67[1.41-1.98]	5.70E-09
PM20D1	1q32	rs11240572	*NUCK1-RAB7L1-SLC41A1*, *PM20D1*	0.74[0.69-0.80]	1.01E-14
STK39	2q24	rs2102808	*STK39*	1.28[1.19-1.38]	1.54E-11
MCCC1/LAMP3	3q27	rs11711441	*MCCC1-LAMP3-MCF2L2*	0.84[0.80-0.89]	8.72E-12
BST1	4p15	rs4698412	*FAM200B-BST1*	0.87[0.83-0.91]	2.28E-10
GAK/DGKQ	4p16	rs1564282	*CPLX1-GAK-TMEM175-DGKQ-SLC26A1*, *IDUA*, *FGFRL1*	1.29[1.20-1.38]	6.54E-13
SNCA	4q21	rs356220	*SNCA-MMRN1*	1.30[1.25-1.34]	3.06E-49
HLA-DRB5	6p21	rs2395163	*HLA-DRB5-HLA-DRB1*, *HLA-DRB6*	0.75[0.68-0.84]	2.90E-07
GPNMB	7p15	rs156429	*GPNMB-MALSU1-IGF2BP3*	0.89[0.86-0.93]	2.69E-10
LRRK2	12q12	rs34778348	*SLC2A13-LRRK2-MUC19*, *CNTN1*	2.23[1.89-2.63]	2.97E-21
CCDC62/HIP1R	12q24	rs12817488	*DENR-HIP1R-VPS37B*, *ABCB9*, *OGFOD2*,	1.17[1.09-1.25]	2.99E-06
MAPT/STH	17q21	H1H2, 900kb inversion	*ARHGAP27*, *PLEKHM1*, *CRHR1*, *SPPL2C-MAPT-STH*, *KANSL1*, *LRRC37A*, *NSFP1*, *ARL17A/B*	0.78[0.75-0.80]	3.54E-52

Chr, chromosomal band; CI, confidence interval. [a]Genes within 100 kb of the most significantly associated SNP annotated from the UCSC genome browser (hg19). Odds ratios and *P*-values are the most significant findings from the PDGene database [150].

An association between PD and *MAPT* and the surrounding 17q21 locus results from an ancient paracentric inversion and was robustly implicated in clinical PD and in autopsy-confirmed series of PD in Caucasian populations [40-42]. Of note, similar variability in the *MAPT* locus has been unequivocally implicated in progressive supranuclear palsy, but not in Alzheimer's disease, although tau is a major component of the neurofibrillary tangle pathology in both conditions [43]. Splice and missense mutations were first described in frontotemporal dementia [44] and the inversion region was subsequently implicated in 17q21.31 microdeletion syndromes [45], but neither pathologic mutations nor tau pathology are found in PD. Thus, genetic variability in neighboring genes, within or flanking the *MAPT* inversion, may contribute. Some examples of additional genes and pathogenic mutations discovered through family-based linkage analysis of parkinsonism are summarized in Table 2.

Genome-wide association studies

While there are several genetic models for disease susceptibility, the 'rare variant common disease' model largely explains Mendelian heritability (one allele of major effect segregating with familial disease) whereas the 'common variant common disease' model (or multiple rare variants on common ancestral haplotypes) forms the theoretical basis of GWAS [46]. Built upon large, multi-institutional consortia, GWAS have dominated the search for novel genes in human traits in recent years. Collectively, more than 2,600 genomic regions (loci) of modest effect size (odds ratio <1.5) have been associated with >350 complex traits and have implicated underlying genes that play a role in disease causality or susceptibility [47]. Genome-wide genotyping, with approximately 0.5 to 1 million markers, is used to assess frequency differences in case-control designs, and is able to capture common genetic contributions to disease in linkage disequilibrium.

The first GWAS for PD were small and underpowered, showing little overlap in results [48,49]. Subsequently, there was an appreciation that larger numbers of markers and subjects are required for meaningful discovery and replication. Past North American and European efforts included analysis of familial parkinsonism [50] and of unrelated case-control series [51-54]. A web-based 'direct- to-consumer' effort, which is based on self report rather than clinical exam, is by far the largest study to date [9]. However, genomic imputation of single nucleotide polymorphisms (SNPs) in linkage disequilibrium allows datasets to be combined. In 2011, the International Parkinson Disease Genomics Consortium conducted a meta-analysis from five GWAS datasets [55]. Over 7 million SNPs based on approximately 1 million genotypes per individual were imputed *in silico*: the data were collected from 5,333 cases and 12,019 controls in the discovery phase, followed by 7,053 cases and 9,007 controls in the replication phase. Six loci previously associated with idiopathic PD were replicated and five new loci were identified (Table 1). Data from the latest mega-meta-GWAS, including over 13,000 patients with PD and 80,000 control subjects, are eagerly awaited.

Table 2 Mendelian mutations in familial parkinsonism

Gene	Mutation(s)	OMIM	Reference(s)
Dominantly inherited, late onset parkinsonism with Lewy pathology			
SNCA	Locus multiplication and missense mutations: A30P, E46K, H50Q, G51N, A53T	*168601*, *605543*	[56,58]
LRRK2	R1437H, R1441H, R1441G, R1441C, Y1699C, G2019S, I2020T	*607060*	[20,31,32]
VPS35	D620N	*614203*	[63,64]
EIF4G1	R1205H	*614251*	[149]
DNAJC13	N855S	*614334*	[93]
Recessively inherited, early-onset or X-linked atypical parkinsonism			
PARK2 (Parkin)	Numerous exon deletions, duplications and missense mutations	*600116*	[109]
PINK1	Rare locus and exon deletions. Numerous missense mutations, including E129X, Q129fsX157, P196L, G309N W437X, G440E, Q456X	*605909*	[147,151]
DJ-1	Deletions and missense: dup168-185, A39S, E64D, D149A, Q163L, L166P, M261I.	*606324*	[123]
DNAJC6	Splice site c.801 -2 A > G and truncating mutation Q734X	*615528*	[70]
ATP13A2	Missense: L552fsX788, M810R, G877R, G1019fsX1021. Small insertions and deletions: 1103insGA, del2742TT	*606693*	[102]
FBXO7	T22M, R378G, R498X	*260300*	[68]
PLA2G6	D331Y, R635Q,R741Q, R747W	*612953*	[152]
ATP6AP2	Splice site mutations		[71]
SYNJ1	Homozygous missense: R258Q	*615530*	[72]

OMIM, Online Mendelian Inheritance in Man, a database that catalogs all the known diseases with a genetic component.

Nevertheless, in the same Caucasian populations, the genetic variance explained is unlikely to increase in proportion to the sample size and investment. While additional loci may be nominated, these assignments are likely to be of smaller effect and will require independent validation.

Overall, the most significant GWAS associations are at chromosomal bands 4q22 and 17q21 and support *SNCA* and *MAPT* assignments, which were first identified in linkage and candidate gene studies [41,56-58]. Nevertheless, GWAS illustrates that the etiology of PD is genetically heterogeneous and novel loci may yet be identified, especially within under-represented non-Caucasian populations; for example, *PARK16*, *BST1* and *SYT11* were identified in a Japanese study [52]. Although a risk allele identified in one population should be a risk allele for all (the genotypic-attributable risk), their frequencies may be ethnically specific and lead to widely divergent population-attributable risks. For example, Caucasians have two major haplotypes (H1 and H2) for the *MAPT* locus; only H1 is present in Asian populations whereas the frequency of H2 is about 20% in Caucasians. The *MAPT* locus is significantly associated with PD in Caucasian studies but does not appear in a Japanese GWAS [51,52]. Conversely, the *PM20D1* locus was most clearly associated with PD in Japanese GWAS (5 to 8% differences between cases and controls) [52], whereas *PM20D1* allele frequencies in Caucasian studies were similar in cases and controls (1% difference) [51].

It is important to note that genomic loci are not disease genes *per se*. Within each genomic locus, there may be numerous genes of which one or all may be candidates contributing to disease risk. An illustration is the *GAK-DGKQ* locus that is significantly associated with PD [9,50,55]. *GAK* and *DGKQ* are in complete linkage disequilibrium and both proteins have important roles in clathrin-mediated vesicular trafficking [59,60]. Furthermore, no coding mutations have been identified in any GWAS locus or gene except for those previously identified through linkage studies in families (that is, *LRRK2* and *SNCA*). Rather, these associations are ascribed to subtle differences in wild-type gene expression, for which RNA silencing and overexpression models may be informative. Common variants of modest effect, on common haplotypes, may lead to modest transcriptional or functional changes. Directly genotyping a 'causal variant' will provide the greatest odds ratio for disease association and provides the rationale for locus-specific sequencing and further association testing. For some disease-associated loci, multiple rare variants of major effect may be responsible, in aggregate, although they will have occurred on the most frequent haplotypes.

Next-generation sequencing for PD

Most of the genetic variants in the human genome are a consequence of mutation with recent population expansion and are present at very low frequencies (<0.5%) [61]. Collectively, rare variants are more common than frequent variants in any given population and each individual has many unique *de novo* point mutations. If these cluster within specific genomic loci in patients with a disease such as PD, they highlight genes or mutational hotspots likely to confer disease susceptibility. To Sanger sequence gene-by-gene in search of a causative variant for PD is a time-consuming and cost-limited effort. Nevertheless, the detection of rare and unique variants via direct sequencing has become more affordable with the advancement of next-generation methods.

Within families, whole-exome sequencing (WES) has proven to be effective in uncovering rare causal mutations of major effect in small sample sizes, and is considerably less expensive than whole-genome sequencing. The first proof-of-concept work discovered pathogenic variants in Freeman-Sheldon syndrome in just four unrelated affected individuals [62]. The first WES study of parkinsonism revealed the p.D620N mutation in *VPS35* (the vacuolar sorting protein 35 gene) by sequencing affected cousin-pairs in autosomal dominant kindreds with late-onset disease [63,64]. These findings have been confirmed worldwide, suggesting that *VPS35* contributes to approximately 1% of familial parkinsonism and 0.2% of sporadic PD [65-67]. WES has become the fastest method for the identification of novel genes in parkinsonism, contributing to the discoveries of *FBXO7*, *WRD45*, *DNAJC6*, *DNAJC13*, *ATP6AP2* and *SYNJ1* in recent years [68-72].

Variants in non-coding and highly conserved genomic regions may also contribute to risk, and might explain the 'missing heritability' underlying complex trait disorders. Whole-genome sequencing (WGS) now enables the sequencing of untranslated regions (UTRs), including gene promoters, enhancers, introns, and 5′ and 3′ UTRs. These contribute to the regulation of gene expression directly through transcription-factor binding, via microRNA and noncoding RNA mechanisms and through alternative exon splicing, and can influence the phenotypic variance of some traits. Two examples relevant to parkinsonism include 5′ UTR expansions in *FMR1* in Fragile X-associated tremor-ataxia syndrome [73] and non-coding mutations in *ATP6AP2* that contribute to aberrant alternative splicing [71].

Next-generation DNA-sequencing panels can be used for novel mutation discovery in a locus-centric approach. In genome-wide family-based studies, when seeking to identify novel genes, it is prudent to examine and exclude those genes already implicated in parkinsonism (including developmental and aging syndromes as described earlier). Unique probes or amplicons can cover exonic regions or span entire loci, and target DNA can be barcoded and sequenced in parallel for multiple individuals. Such panels have many benefits: first, they provide affordable and rapid genetic diagnoses to better inform patient treatment, without problems arising from incidental findings [74]; second, they allow the discovery of as-yet-unknown mutations in genes known to be related to PD; and third, in aggregate, the results nominate genetic variants that may further explain the heterogeneity of clinical and pathologic presentations. Similar bioinformatic filters might be applied to WES or WGS *in silico*. In a gene-centric approach, with a limited number of genes studied per assay relative to WES or WGS, the interpretation is simplified. Nevertheless, custom capture or amplification methods can produce artifacts as oligomers may not be perfectly complementary to the human genome being interrogated, and this may introduce some allelic bias.

Another pitfall of next-generation sequencing (NGS) methods is the inability to sequence repetitive regions and structural variations. For example, *SNCA* multiplications, the *GBA* loci versus its pseudogene, or indeed any repetitive regions may confound sequence analysis; most sequence read lengths are relatively short (approximately 100 to 200 bp) and may be misaligned as a consequence. Almost half of the human genome consists of repeats that play an active role in genome evolution, although large structural rearrangements may result in disease [75]. Inadequate sequence read-depth may also lead to annotation errors. Most publically shared WGS has been performed at relatively low read depths; for example, the 1000 Genomes effort achieved 2x to 6x coverage (www.1000genomes.org). For WES, >100x coverage is considered necessary for diagnostic testing [76]. Nevertheless, specific NGS results should be validated by Sanger sequencing to prevent false positives. Higher-coverage sequencing, longer read lengths and innovative bioinformatic approaches continue to offer significant improvements.

In studies involving WES or WGS of familial PD, rare variants can be prioritized for validation and replication in additional samples by looking at the intersection of those shared by affected family members and not shared by elderly unaffected relatives. Typically, in approximately 58 Mb of exome sequencing approximately 80,000 single nucleotide variants are observed per individual, of which approximately 250 encode substitutions not annotated in public databases. To further reduce this number, a powerful approach has been to compare affected cousin-pairs, in contrast to affected sibling-pairs, to reduce the number of alleles shared by descent (from approximately 125 to approximately 31 variants). In 'pair-wise sharing analysis' non-synonymous missense mutations that have a frequency less than the incidence of PD (<0.003), which are evolutionarily conserved and predicted as damaging, are prioritized as good candidates for follow-up. However, caution is warranted in comparative analyses: the phenocopy rate of late-onset PD is approximately 18% in families in which a monogenic cause of disease is already defined [77]. The strategy may also exclude causal or risk variants. For example, *SNCA* p.A53T was the first mutation identified in PD but is neither evolutionarily conserved nor predicted to be deleterious [56].

Molecular pathways in PD

Many genetic components now appear to contribute to the pathogenesis of parkinsonism. It is largely unknown whether the proteins involved function in overlapping biological pathways, whose dysfunction results in the progressive loss of striatal dopaminergic innervation and death of midbrain nigral neurons. Nevertheless, some relationships, involving the perturbation of relatively few cellular systems, are apparent (Figure 1). The affected systems include synaptic transmission, endosomal trafficking, lysosomal-autophagy and energy metabolism or mitophagy.

Synaptic transmission

Alpha-synuclein protein is found abundantly at the presynaptic terminals of neurons and is involved in synaptic release [13] (Figure 1). Monomeric forms of alpha-synuclein may contribute to endophilin-A1-related membrane curvature, facilitating both synaptic vesicle exo- and endocytosis [78]. A tetrameric conformation has been proposed for alpha-synuclein and this might provide a parsimonious explanation of how amino-terminal point mutations lead to the same functional deficits [79,80]. The conformation of alpha-synuclein has previously revealed physiologic interactions with mitochondria [81] and with presynaptic tubulovesicular or endosomal structures when alpha-synuclein is overexpressed in transgenic mice [82].

At the synapse, LRRK2 levels regulate glutamate transmission, dopamine-dependent plasticity and striatal signal transduction [83,84]. LRRK2 protein levels and mutant-specific phenotypes have long been observed in neuritic outgrowth (affecting branching and length) in primary cultures [85,86] and in neurogenesis *in vivo* [87]. In *Drosophila*, LRRK2 kinase has been shown to regulate the EndoA phosphorylation cycle, and pathogenic mutations appear to impede synaptic endocytosis [88] (Figure 1). LRRK2 protein is also reported to interact with the dynamin superfamily of GTPases, which mediate both membrane scission in clathrin-induced endocytosis and mitochondrial fission and fusion [89]. In *Caenorhabditis elegans*, knockout of LRK-1 (the single homolog of mammalian *LRRK1* and *LRRK2*) leads to impairment in presynaptic protein sorting and axonal trafficking [90]. Several important functions have also been ascribed to the LRRK2 protein complex in non-neuronal cells, including kidney cells [26,91], and in innate immunity [92].

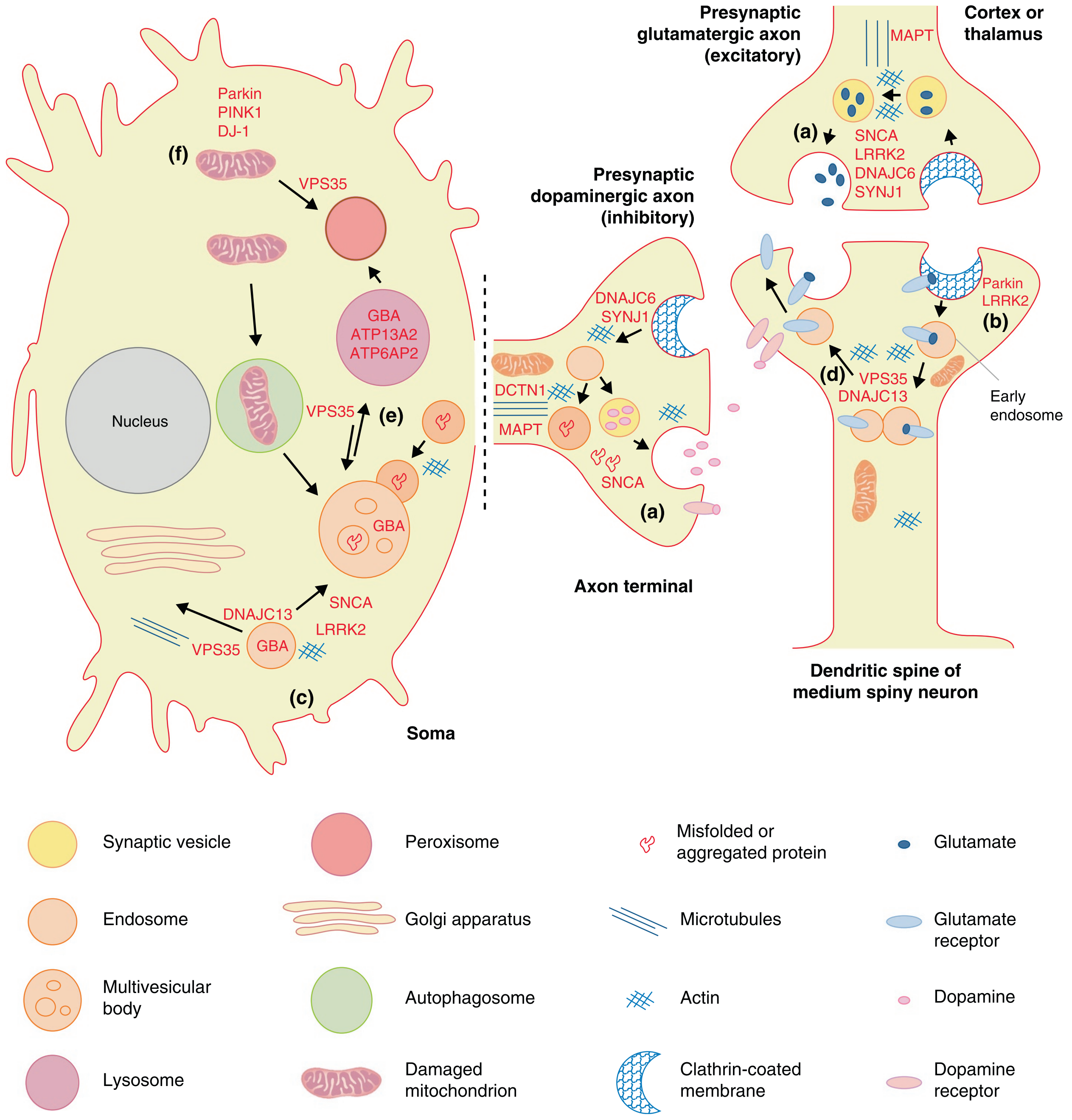

Parkin
PINK1
DJ-1
(f)
VPS35
Nucleus
GBA
ATP13A2
ATP6AP2
VPS35
(e)
GBA
DNAJC13
SNCA
LRRK2
VPS35
GBA
(c)
Soma
Presynaptic dopaminergic axon (inhibitory)
DNAJC6
SYNJ1
DCTN1
MAPT
SNCA
(a)
Axon terminal
Presynaptic glutamatergic axon (excitatory)
Cortex or thalamus
MAPT
(a)
SNCA
LRRK2
DNAJC6
SYNJ1
Parkin
LRRK2
(b)
(d)
VPS35
DNAJC13
Early endosome
Dendritic spine of medium spiny neuron
Synaptic vesicle
Peroxisome
Misfolded or aggregated protein
Glutamate
Endosome
Golgi apparatus
Microtubules
Glutamate receptor
Multivesicular body
Autophagosome
Actin
Dopamine
Lysosome
Damaged mitochondrion
Clathrin-coated membrane
Dopamine receptor

Figure 1 (See legend on next page.)

(See figure on previous page.)
Figure 1 Cellular processes implicated in familial late-onset, Lewy body parkinsonism. In late onset parkinsonism, which is most reminiscent of Lewy body PD, a novel synthesis is emerging whereby regulatory steps in synaptic neurotransmission, receptor recycling, endosomal trafficking and lysosomal degradation are controlled by relatively few proteins. Studies in neurons and brain are limited but many reciprocal connections are apparent [8]. For example, **(a)** SNCA functions with heat shock chaperone Hsc70 and SNAP25 to promote membrane SNARE complex assembly and exocytosis in neurotransmission. **(b)** In mammalian cells, LRRK2 may also regulate cleavage of invaginated endocytic membrane through interaction with dynamin. Several other genes, including *DNAJC6* (with homology to *GAK*) and *SYNJ1*, encode proteins important in clathrin uncoating. In *Drosophila*, LRK1 (the homolog of mammalian LRRK1 and LRRK2) phosphorylates endophilin A to directly regulate endocytosis. In the mammalian striatum the expression levels of SNCA and endophilin A are reciprocally related. **(c)** In the endosome, VPS35 and the retromer cargo-selective-complex (CSC; comprising VPS35-VPS26 and VPS29) play an important role in membrane protein cargo sorting. The CSC is best described in endosome to trans-Golgi network retrieval in the soma, most specifically in recycling cation-independent mannose-6-phosphate receptor that traffics acidic hydrolases (including GBA) to the lysosome. However, recent studies in neurons have revealed that VPS35 and WASH are also crucial in protein recycling **(d)** of specific synaptic receptors from early endosome to plasma membrane, and **(e)** in the V-ATPase required for lysosomal acidification. Importantly, CSC trafficking can be mediated by sorting nexin interaction with dynactin p150$^{\text{Glued}}$ (DCTN1), dynein, tau (MAPT) and microtubules, or via the WASH complex, DNAJC13 and actin polymerization. DNAJC13 was first described in endocytosis in *C. elegans*, rather than in post-endocytic trafficking, and like SNCA requires Hsc70 to function. LRRK2 has also been shown to interact with VPS35 and microtubules, and like SNCA may be intrinsically targeted to lysosomal membranes by protein motifs for chaperone-mediated autophagy. Clearance of insoluble SNCA aggregates is also mediated by the endosomal system and lysosomal degradation [153]. Several other lysosomal proteins, including ATP13A2 and ATP6AP2, have been implicated in atypical parkinsonism. EIF4G1, through mTOR regulation of protein translation, serves to balance autophagic activity and metabolism or ATP levels, whereas **(f)** PINK1 and Parkin are intrinsically involved in mitochondrial quality control in early onset parknsonism.

Recessively inherited mutations in *DNAJC6* have recently been identified in juvenile parkinsonism [70]. *DNAJC6* encodes auxilin, a homolog of cyclin-G associated kinase (GAK; Table 1), which is preferentially expressed in neurons and involved in clathrin uncoating and synaptic vesicle recycling. Similarly, recessively inherited mutations in *SYNJ1*, encoding synaptojamin, that complexes with Hsc70 and auxilin, have been implicated in disease [72].

Endosomal trafficking

Endosomal trafficking is a highly complex and dynamic cellular process whereby vesicles or cargos that are internalized at the plasma membrane are subsequently recycled, directly or via the trans-Golgi network, and targeted for degradation by lysosomal autophagy. Mutations in VPS35 and RME-8 (receptor-mediated endocytosis 8, also known as DNAJC13) were recently linked to late-onset Lewy body parkinsonism and directly implicate endosomal trafficking in disease pathogenesis [63,93]. Neurons have a critical need to recycle membrane receptors. This can be accomplished through the clathrin-independent retromer system, a tubulovesicular tripartite complex of VPS26 (vacuolar sorting protein 26), VPS29 and VPS35 that relies on sorting nexins to stipulate the destinations of specific cargos, such as neurotransmitter receptors. Multiple VPS35 subunits coalesce about FAM21, a subunit of the WASH (Wiskott-Aldrich syndrome protein and scar homolog) complex, to mediate dynamic actin remodeling [94]. RME-8 also binds sorting nexins and FAM21 to influence WASH and cargo trafficking [95]. VPS35 may also physically interact with LRRK2 and Rab7L1 (within the PM20D1 locus; Table 1) to influence these processes [96].

Lysosomal autophagy

Lysosomes have an essential function in maintaining protein and organelle integrity within cells and impaired lysosomal function may play an important role in the pathogenesis of PD. Aggregated alpha-synuclein, in the form of Lewy neuritic or Lewy body inclusions, that fails to be degraded by proteosomal or lysosomal systems is the pathologic hallmark of PD. It is presently unknown whether the intracellular protein aggregation observed in most late-onset neurodegenerative diseases is a cause or consequence of dysfunction in these pathways [97]. The formation of intracellular aggregated alpha-synuclein or tau inclusions, albeit not a primary pathology, is also found in several ceroid lipofuscinosis disorders. These include glycolipid storage diseases such as Gaucher disease and Niemann-Pick type C that are most prevalent in Ashkenazi Jewish communities. Although the *GBA* (glucocerebrosidase gene) mutations are recessively inherited in Gaucher disease, heterozygote carriers have an increased prevalence of PD and dementia with Lewy bodies (Table 1) [98]. A lysosomal pathway for parkinsonism centered around ceramide metabolism has been hypothesized [99]. Loss of GBA activity increases intracellular glucosylceramide accumulation, resulting in decreased lysosomal degradation and subsequent accumulation of alpha-synuclein [100]. Whether the latter reflects impaired GBA trafficking from the endoplasmic reticulum and Golgi to lysosomes or whether it results directly from lysosomal dysfunction is unclear. In genetically engineered mice, *GBA* mutations promote alpha-synuclein accumulation in a dose- and time-dependent manner, with the animals developing Lewy-like pathology in the brain and associated motor and cognitive phenotypes. By contrast, loss of GBA activity results in neuronal ubiquitinopathy and formation of axonal spheroids, a phenotype that is shared with other lysosomal storage disorders prior to increased alpha-synuclein concentrations [101].

Two juvenile or early-onset forms of atypical parkinsonism result directly from mutations in lysosomal proteins. X-linked parkinsonism, with onset in males between 14 and 50 years of age and associated with post-mortem tau pathology, is a consequence of splicing or protein isoform deficits in *ATP6AP2* (encoding ATPase, H^+ transporting, lysosomal accessory protein 2) [71]. Recessively inherited p*athogenic mutations in* ATP13A2 (ATPase type 13A2 gene) also result in impaired lysosomal proteolysis, leading to Kufor-Rakeb syndrome [102]. Patients and knock-out mice develop ceroid lipofuscin neuronal pathology, and the mice show concomitant upregulation of alpha-synuclein protein in the hippocampus [103]. Many genes are implicated in neurodegeneration with brain iron accumulation, including *ATP13A2*, *PLA2G6*, *PANK2*, *C19orf12*, *FA2H*, *WDR45*, *FTL*, *CP*, and *DCAF17* [104].

Mitochondrial metabolism

The earliest link between mitochondrial dysfunction and parkinsonism was observed in illicit drug users. MPTP (1-methyi-4-phenyl-1,2,3,6-tetrahydropyridine) is specifically transported into dopamine neurons via the dopamine transporter and is then oxidized into toxic MPP^+, a non-competitive complex inhibitor of the electron transport chain [105]. Deficits of mitochondrial complex I have been noted in idiopathic PD [106], although evidence from direct sequencing studies of normal brain has proven equivocal. Mitochondrial mutations in humans lead to several neuromuscular disorders [107]. While the majority are not associated with parkinsonism, similar movement disorders with or without chronic progressive ophthalmoplegia can be caused by mutations in mitochondrial DNA polymerase γ (POLG), a proofreading enzyme. Mouse models with defective POLG exhibit premature ageing whereas older homozygous, but not heterozygous, POLG mice show significant reductions in striatal dopaminergic terminals as well as deficits in motor function [108].

The importance of mitochondria in parkinsonism is highlighted by the identification of mutations in several genes within a common pathway for mitophagy. Mutations in the *PARK2* (*parkin*) gene result in a recessive form of early-onset parkinsonism [109]. Parkin protein was first described as a proteosomal E3 ubiquitin ligase responsible for K48 substrate polyubiquination (targeting to the proteosome) and K63 monoubiquination (for signaling) [110]. In addition, parkin may have several physiological roles in neurons - for example, in Eps15 monoubiquination [111] and in the regulation of neuronal apoptosis as part of a SCF-like complex [112]. Most highlighted is the role of parkin in regulating the degradation of depolarized or uncoupled mitochondria, in concert with PINK1 (Pten-induced kinase 1) and FBXO7 (F-box domain-containing protein), which are also genes implicated in recessive early-onset parkinsonism [68,113]. *Drosophila* parkin and *PINK1* knockout models exhibit similar mitochondrial and wing phenotypes, and a series of elegant experiments has demonstrated that PINK1 is required for the recruitment of parkin to mitochondria [114,115]. The crystal structure of parkin has now been solved [116] and PINK1 has been shown to phosphorylate ubiquitin required for parkin's activation [117]. Two recent RNA interference screens have identified upstream regulators of mitophagy, albeit with limited overlap [118,119]. These include TOMM7, essential for stabilizing PINK1 on the outer mitochondrial membrane; HSPA1L and BAG4, which may help to regulate parkin translocation to mitochondria; and SIAH3, which is localized to mitochondria and inhibits PINK1 after mitochondrial damage, thereby reducing parkin translocation. Hexokinase activity, occurring downstream of Akt but upstream of PINK1, may also be required in the recruitment of parkin to depolarized mitochondria [119]. STOML2, mitofusin1/2, GRP75, HSP60, LRPPRC, and TUFM have been nominated as downstream targets of the PINK1/parkin pathway [120-122]. DJ-1 mutations, which result in early-onset parkinsonism [123], may also regulate PINK1-dependent parkin translocation to depolarized mitochondria [124]. DJ-1 deficiency leads to altered mitochondrial morphology and increased levels of reactive oxygen species (ROS) [124]. Traditionally, knockout mouse models of parkin, PINK1 or DJ-1 result in mitochondrial dysfunction [125] but do not develop the locomotor phenotype of parkinsonism, nigral neuronal loss or Lewy-body pathology; rather they have elevated dopaminergic tone due to deficits in D2 presynaptic regulation of release [126,127]. However, a recent conditional parkin knockout mouse model demonstrated progressive loss of dopamine neurons in a PARIS-dependent pathway [128]. Thus, protein components of the parkin/PINK1 mitochondrial pathways remain plausible therapeutic targets for human carriers of these mutations and potentially for idiopathic PD.

While mitochondrial mutations and proteins involved in mitophagy have been directly implicated in parkinsonism, there is accumulating albeit indirect evidence that mitochondrial function is central to disease pathogenesis and/or progression. For example, alpha-synuclein overexpression may impair mitochondrial activity, thereby accumulating mitochondrial DNA damage and degeneration, ultimately resulting in neuronal death [129]. For LRRK2, mitochondrial pathology was observed in human dopaminergic neurons derived from inducible pluripotent cell lines of *LRRK2* p.G2019S carriers [130] and in aging p.G2019S transgenic mice [131]. In primary cortical neurons, overexpression of wild-type LRRK2 and of pathogenic mutant LRRK2 proteins both increased recruitment of mitochondrial dynamin-like protein to fragmenting mitochondria [132].

Conclusions and future directions

Idiopathic PD, albeit a sporadic disorder with low heritability, now appears to have a significant genetic component. Many genes have been identified and several more will probably emerge from a variety of complementary approaches. Most immediately, targeted genomic sequencing might identify functional variants in GWAS-associated loci, whereas additional GWAS in non-Caucasian populations might identify novel loci. WES in families with parkinsonism is also a pragmatic step to identify more genes using a concordant 'pairwise' approach. Ultimately, collaborative pooled analysis of exome data might facilitate association analysis; power estimates suggest that as few as 50 exomes may be sufficient to identify a novel locus for recessive disease, although 10,000 are likely to be required to identify a novel dominant gene mutation [133]. Nevertheless, segregation of rare mutations with disease and/or functional studies will be needed. Inevitably, with less penetrant variants, the results from these approaches will become increasingly difficult to interpret. Hence, rather than focusing on PD as the trait, more investment may be warranted on longitudinal studies. In addition, better characterization of trait components, such as cognition in PD, symptom progression and response to medication, would enable further genetic variability to be identified. Similar analyses in more genetically homogeneous populations (employing linkage, association and genome sequencing) and in sample series or pedigrees of sufficient size and structure may enable the joint contribution of genes and environment to be assessed meaningfully. We suggest that a genetic predisposition to PD should not be considered 'causal', rather disease reflects chronic molecular dysfunction and the failure of age-associated compensation.

To date, four biological pathways have been implicated in familial parkinsonism: synaptic neurotransmission, endosomal trafficking, lysosomal autophagy and mitochondrial metabolism. Direct interactions between genetic components and these pathways are emerging, whether *VPS35* and *RME-8* in late onset PD, or *parkin* and *PINK1* in mitophagy in early onset parkinsonism. Although the processes highlighted may be viewed separately, several employ the same protein machinery and may be temporally and functionally related. For example, synaptic dysfunction, resulting from or leading to alpha-synucleinopathy, impairs the balance of exo- and endocytosis, neurotransmission and early endosomal receptor recycling. These changes will alter flux through the endosomal pathway and ultimately place demand on autophagy and lysosomal fusion with multivesicular bodies. Many of the same proteins are involved in more than one of the four pathways; for example, VPS35 directly affects both early-endosome receptor recycling in dendritic spines and lysosomal ATPase recycling from multivesicular bodies [134]. Similarly, LRRK2 appears to be centrally involved in neurite outgrowth and in membrane protein cargo sorting and trafficking, interacts with dynamin GTPases, and may regulate endophilin phosphorylation, membrane scission and endocytosis. VPS35, RME-8 and potentially LRRK2 coordinate the WASH complex in specific actin networks underlying membrane deformation, tubulation and cargo trafficking. The caveat is that many of the biological insights may be model specific, or have yet to be performed in vertebrate systems in post-mitotic neurons or in brain.

In conclusion, genomic and genetic investigations should continue to be a main priority for future research. Knowledge of the pathogenic pathways underlying the etiology and ontology of PD clearly facilitates an understanding of common protein components and central processes that are crucial for therapeutic development. Our understanding of the normal physiology of the brain, of specific neuronal populations, protein pathways and the function of individual proteins, is rudimentary. When it comes to formulating a molecular synthesis of the pathways involved in parkinsonism, genetic insights may be unbiased and unequivocal but those insights must be carefully weighted. Genetic association is far from causation and much work is required to understand the specific contribution of GWAS, let alone to translate the information from such studies into novel treatments to slow or halt the progression of idiopathic PD. In interpreting linkage and exome studies, the phenotypes of patients and families must be carefully considered. The heterogeneity of parkinsonism is considerable and there are many forms that may not be etiologically related. Arguably, findings from familial late-onset Lewy-body parkinsonism may be more relevant to idiopathic PD than those from atypical and/or early-onset forms. Brain pathology, long required for a definite diagnosis, appears increasingly pleomorphic in genetically defined disease, even for the same mutation in the same family. Such alternative pathologic outcomes become more intriguing as specific molecular deficits in membrane protein sorting and cargo trafficking are revealed. That aggregate alpha-synuclein and tau pathologies may be seeded and transmissible throughout the cerebrum presents an attractive means to explore Braak staging (a regional and temporal scheme for the progression of these inclusion body pathologies) and vulnerable cell populations in specific genetic backgrounds. Much neuroscience in PD was derived from model systems based on toxin administration, and may not accurately reflect the human condition. Why the *substantia nigra pars compacta* is selectively lost in PD remains enigmatic, but through human genetics we now have relevant molecular targets and tools to investigate this.

With such advances, therapeutic prospects for disease modification (neuroprotection) should be viewed with more optimism [135].

Abbreviations

ATP13A2: ATPase type 13A2 gene; ATP6AP2: ATPase, H^+ transporting, lysosomal accessory protein 2; CSC: cargo-selective-complex; DBS: deep brain stimulation; EIF4G1: elongation initiation factor 4 gamma 1; FBXO7: F-box domain-containing protein 7; FTDP-17: frontotemporal dementia with parkinsonism linked to chromosome 17; GAK: cyclin-G associated kinase; *GBA*: glucocerebrosidease gene; GWAS: genome-wide association study; H&E staining: hematoxylin and eosin staining; LRRK2: leucine-rich repeat kinase 2 protein; MAPT: microtubule-associated protein tau; MPTP: 1-methyi-4-phenyl-1,2,3,6-tetrahydropyridine; NACP: non-amyloid component of plaques; NGS: next-generation sequencing; OMIM: Online Mendelian Inheritance in Man; PD: Parkinson's disease; POLG: DNA polymerase γ; RME-8: receptor-mediated endocytosis 8; ROS: reactive oxygen species; SN: *substantial nigra pars compacta*; *SNCA*: alpha-synuclein gene; SNP: single nucleotide polymorphism; UCHL1: Ubiquitin C-terminal hydrolase1; UTR: untranslated region; VPS35: vacuolar sorting protein 35; WASH: Wiskott-Aldrich syndrome protein and scar homolog; WES: whole-exome sequencing; WGS: whole-genome sequencing.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

We are indebted to the Canada Excellence Research Chairs program and the Dr Donald Rix BC Leadership Chair in Genetic Medicine for ongoing support of the Centre of Applied Neurogenetics and Translational Neuroscience.

Published: 30 June 2014

References

1. de Lau LM, Breteler MM: **Epidemiology of Parkinson's disease.** *Lancet Neurol* 2006, **5:**525–535.
2. Fahn S: **Description of Parkinson's disease as a clinical syndrome.** *Ann N Y Acad Sci* 2003, **991:**1–14.
3. Benabid AL, Chabardes S, Mitrofanis J, Pollak P: **Deep brain stimulation of the subthalamic nucleus for the treatment of Parkinson's disease.** *Lancet Neurol* 2009, **8:**67–81.
4. Langston JW: **The Parkinson's complex: parkinsonism is just the tip of the iceberg.** *Ann Neurol* 2006, **59:**591–596.
5. Aarsland D, Kurz MW: **The epidemiology of dementia associated with Parkinson's disease.** *Brain Pathol* 2010, **20:**633–639.
6. Hughes AJ, Daniel SE, Ben-Shlomo Y, Lees AJ: **The accuracy of diagnosis of parkinsonian syndromes in a specialist movement disorder service.** *Brain* 2002, **125:**861–870.
7. Goedert M, Spillantini MG, Del Tredici K, Braak H: **100 years of Lewy pathology.** *Nat Rev Neurol* 2013, **9:**13–24.
8. Trinh J, Farrer M: **Advances in the genetics of Parkinson disease.** *Nat Rev Neurol* 2013, **9:**445–454.
9. Do CB, Tung JY, Dorfman E, Kiefer AK, Drabant EM, Francke U, Mountain JL, Goldman SM, Tanner CM, Langston JW, Wojcicki A, Eriksson N: **Web-based genome-wide association study identifies two novel loci and a substantial genetic component for Parkinson's disease.** *PLoS Genet* 2011, **7:**e1002141.
10. Wirdefeldt K, Gatz M, Reynolds CA, Prescott CA, Pedersen NL: **Heritability of Parkinson disease in Swedish twins: a longitudinal study.** *Neurobiol Aging* 1923, **2011:**e1–e8.
11. Ross OA, Braithwaite AT, Skipper LM, Kachergus J, Hulihan MM, Middleton FA, Nishioka K, Fuchs J, Gasser T, Maraganore DM, Adler CH, Larvor L, Chartier-Harlin MC, Nilsson C, Langston JW, Gwinn K, Hattori N, Farrer MJ: **Genomic investigation of alpha-synuclein multiplication and parkinsonism.** *Ann Neurol* 2008, **63:**743–750.
12. Abeliovich A, Schmitz Y, Farinas I, Choi-Lundberg D, Ho WH, Castillo PE, Shinsky N, Verdugo JM, Armanini M, Ryan A, Hynes M, Phillips H, Sulzer D, Rosenthal A: **Mice lacking alpha-synuclein display functional deficits in the nigrostriatal dopamine system.** *Neuron* 2000, **25:**239–252.
13. Burre J, Sharma M, Tsetsenis T, Buchman V, Etherton MR, Sudhof TC: **Alpha-synuclein promotes SNARE-complex assembly in vivo and in vitro.** *Science* 2010, **329:**1663–1667.
14. Olanow CW, Prusiner SB: **Is Parkinson's disease a prion disorder?** *Proc Natl Acad Sci U S A* 2009, **106:**12571–12572.
15. Kordower JH, Chu Y, Hauser RA, Freeman TB, Olanow CW: **Lewy body-like pathology in long-term embryonic nigral transplants in Parkinson's disease.** *Nat Med* 2008, **14:**504–506.
16. Luk KC, Kehm V, Carroll J, Zhang B, O'Brien P, Trojanowski JQ, Lee VM: **Pathological alpha-synuclein transmission initiates Parkinson-like neurodegeneration in nontransgenic mice.** *Science* 2012, **338:**949–953.
17. McCormack AL, Mak SK, Henderson JM, Bumcrot D, Farrer MJ, Di Monte DA: **Alpha-synuclein suppression by targeted small interfering RNA in the primate substantia nigra.** *PLoS One* 2010, **5:**e12122.
18. Valera E, Masliah E: **Immunotherapy for neurodegenerative diseases: focus on alpha-synucleinopathies.** *Pharmacol Ther* 2013, **138:**311–322.
19. Trinh J, Amouri R, Duda JE, Morley JF, Read M, Donald A, Vilarino-Guell C, Thompson C, Szu Tu C, Gustavsson EK, Ben Sassi S, Hentati E, Zouari M, Farhat E, Nabli F, Hentati F, Farrer MJ: **A comparative study of Parkinson's disease and leucine-rich repeat kinase 2 p.G2019S parkinsonism.** *Neurobiol Aging* 2014, **35:**1125–1131.
20. Ozelius LJ, Senthil G, Saunders-Pullman R, Ohmann E, Deligtisch A, Tagliati M, Hunt AL, Klein C, Henick B, Hailpern SM, Lipton RB, Soto-Valencia J, Risch N, Bressman SB: **LRRK2 G2019S as a cause of Parkinson's disease in Ashkenazi Jews.** *N Engl J Med* 2006, **354:**424–425.
21. Clark LN, Wang Y, Karlins E, Saito L, Mejia-Santana H, Harris J, Louis ED, Cote LJ, Andrews H, Fahn S, Waters C, Ford B, Frucht S, Ottman R, Marder K: **Frequency of LRRK2 mutations in early- and late-onset Parkinson disease.** *Neurology* 2006, **67:**1786–1791.
22. Hentati F, Trinh J, Thompson C, Nosova E, Farrer MJ, Aasly JO: **LRRK2 parkinsonism in Tunisia and Norway: a comparative study of disease penetrance.** *Neurology* 2014, **83:**1–2.
23. Hulihan MM, Ishihara-Paul L, Kachergus J, Warren L, Amouri R, Elango R, Prinjha RK, Upmanyu R, Kefi M, Zouari M, Sassi SB, Yahmed SB, El Euch-Fayeche G, Matthews PM, Middleton LT, Gibson RA, Hentati F, Farrer MJ: **LRRK2 Gly2019Ser penetrance in Arab-Berber patients from Tunisia: a case-control genetic study.** *Lancet Neurol* 2008, **7:**591–594.
24. Thaler A, Ash E, Gan-Or Z, Orr-Urtreger A, Giladi N: **The LRRK2 G2019S mutation as the cause of Parkinson's disease in Ashkenazi Jews.** *J Neural Transm* 2009, **116:**1473–1482.
25. Gardet A, Benita Y, Li C, Sands BE, Ballester I, Stevens C, Korzenik JR, Rioux JD, Daly MJ, Xavier RJ, Podolsky DK: **LRRK2 is involved in the IFN-gamma response and host response to pathogens.** *J Immunol* 2010, **185:**5577–5585.
26. Herzig MC, Kolly C, Persohn E, Theil D, Schweizer T, Hafner T, Stemmelen C, Troxler TJ, Schmid P, Danner S, Schnell CR, Mueller M, Kinzel B, Grevot A, Bolognani F, Stirn M, Kuhn RR, Kaupmann K, van der Putten PH, Rovelli G, Shimshek DR: **LRRK2 protein levels are determined by kinase function and are crucial for kidney and lung homeostasis in mice.** *Hum Mol Genet* 2011, **20:**4209–4223.
27. Umeno J, Asano K, Matsushita T, Matsumoto T, Kiyohara Y, Iida M, Nakamura Y, Kamatani N, Kubo M: **Meta-analysis of published studies identified eight additional common susceptibility loci for Crohn's disease and ulcerative colitis.** *Inflamm Bowel Dis* 2011, **17:**2407–2415.
28. Funayama M, Li Y, Tomiyama H, Yoshino H, Imamichi Y, Yamamoto M, Murata M, Toda T, Mizuno Y, Hattori N: **Leucine-rich repeat kinase 2 G2385R variant is a risk factor for Parkinson disease in Asian population.** *Neuroreport* 2007, **18:**273–275.
29. Ross OA, Toft M, Whittle AJ, Johnson JL, Papapetropoulos S, Mash DC, Litvan I, Gordon MF, Wszolek ZK, Farrer MJ, Dickson DW: **Lrrk2 and Lewy body disease.** *Ann Neurol* 2006, **59:**388–393.
30. Rajput A, Dickson DW, Robinson CA, Ross OA, Dachsel JC, Lincoln SJ, Cobb SA, Rajput ML, Farrer MJ: **Parkinsonism, Lrrk2 G2019S, and tau neuropathology.** *Neurology* 2006, **67:**1506–1508.
31. Ujiie S, Hatano T, Kubo S, Imai S, Sato S, Uchihara T, Yagishita S, Hasegawa K, Kowa H, Sakai F, Hattori N: **LRRK2 I2020T mutation is associated with tau pathology.** *Parkinsonism Relat Disord* 2012, **18:**819–823.
32. Zimprich A, Biskup S, Leitner P, Lichtner P, Farrer M, Lincoln S, Kachergus J, Hulihan M, Uitti RJ, Calne DB, Stoessl AJ, Pfeiffer RF, Patenge N, Carbajal IC, Vieregge P, Asmus F, Müller-Myhsok B, Dickson DW, Meitinger T, Strom TM, Wszolek ZK, Gasser T: **Mutations in LRRK2 cause autosomal-dominant parkinsonism with pleomorphic pathology.** *Neuron* 2004, **44:**601–607.

33. Marin I, van Egmond WN, van Haastert PJ: **The Roco protein family: a functional perspective.** *FASEB J* 2008, **22:**3103–3110.
34. Ross OA, Soto-Ortolaza AI, Heckman MG, Aasly JO, Abahuni N, Annesi G, Bacon JA, Bardien S, Bozi M, Brice A, Brighina L, Van Broeckhoven C, Carr J, Chartier-Harlin MC, Dardiotis E, Dickson DW, Diehl NN, Elbaz A, Ferrarese C, Ferraris A, Fiske B, Gibson JM, Gibson R, Hadjigeorgiou GM, Hattori N, Ioannidis JP, Jasinska-Myga B, Jeon BS, Kim YJ, Klein C, *et al*: **Association of LRRK2 exonic variants with susceptibility to Parkinson's disease: a case-control study.** *Lancet Neurol* 2011, **10:**898–908.
35. West AB, Moore DJ, Biskup S, Bugayenko A, Smith WW, Ross CA, Dawson VL, Dawson TM: **Parkinson's disease-associated mutations in leucine-rich repeat kinase 2 augment kinase activity.** *Proc Natl Acad Sci U S A* 2005, **102:**16842–16847.
36. Jaleel M, Nichols RJ, Deak M, Campbell DG, Gillardon F, Knebel A, Alessi DR: **LRRK2 phosphorylates moesin at threonine-558: characterization of how Parkinson's disease mutants affect kinase activity.** *Biochem J* 2007, **405:**307–317.
37. Dzamko N, Deak M, Hentati F, Reith AD, Prescott AR, Alessi DR, Nichols RJ: **Inhibition of LRRK2 kinase activity leads to dephosphorylation of Ser(910)/Ser(935), disruption of 14-3-3 binding and altered cytoplasmic localization.** *Biochem J* 2010, **430:**405–413.
38. Deng X, Dzamko N, Prescott A, Davies P, Liu Q, Yang Q, Lee JD, Patricelli MP, Nomanbhoy TK, Alessi DR, Gray NS: **Characterization of a selective inhibitor of the Parkinson's disease kinase LRRK2.** *Nat Chem Biol* 2011, **7:**203–205.
39. Gilsbach BK, Kortholt A: **Structural biology of the LRRK2 GTPase and kinase domains: implications for regulation.** *Front Mol Neurosci* 2014, **7:**32.
40. de Silva R, Hardy J, Crook J, Khan N, Graham EA, Morris CM, Wood NW, Lees AJ: **The tau locus is not significantly associated with pathologically confirmed sporadic Parkinson's disease.** *Neurosci Lett* 2002, **330:**201–203.
41. Skipper L, Wilkes K, Toft M, Baker M, Lincoln S, Hulihan M, Ross OA, Hutton M, Aasly J, Farrer M: **Linkage disequilibrium and association of MAPT H1 in Parkinson disease.** *Am J Hum Genet* 2004, **75:**669–677.
42. Zody MC, Jiang Z, Fung HC, Antonacci F, Hillier LW, Cardone MF, Graves TA, Kidd JM, Cheng Z, Abouelleil A, Chen L, Wallis J, Glasscock J, Wilson RK, Reily AD, Duckworth J, Ventura M, Hardy J, Warren WC, Eichler EE: **Evolutionary toggling of the MAPT 17q21.31 inversion region.** *Nat Genet* 2008, **40:**1076–1083.
43. Vandrovcova J, Anaya F, Kay V, Lees A, Hardy J, de Silva R: **Disentangling the role of the tau gene locus in sporadic tauopathies.** *Curr Alzheimer Res* 2010, **7:**726–734.
44. Hutton M, Lendon CL, Rizzu P, Baker M, Froelich S, Houlden H, Pickering-Brown S, Chakraverty S, Isaacs A, Grover A, Hackett J, Adamson J, Lincoln S, Dickson D, Davies P, Petersen RC, Stevens M, de Graaff E, Wauters E, van Baren J, Hillebrand M, Joosse M, Kwon JM, Nowotny P, Che LK, Norton J, Morris JC, Reed LA, Trojanowski J, Basun H, *et al*: **Association of missense and 5′-splice-site mutations in tau with the inherited dementia FTDP-17.** *Nature* 1998, **393:**702–705.
45. Koolen DA, Sharp AJ, Hurst JA, Firth HV, Knight SJ, Goldenberg A, Saugier-Veber P, Pfundt R, Vissers LE, Destrée A, Grisart B, Rooms L, Van der Aa N, Field M, Hackett A, Bell K, Nowaczyk MJ, Mancini GM, Poddighe PJ, Schwartz CE, Rossi E, De Gregori M, Antonacci-Fulton LL, McLellan MD 2nd, Garrett JM, Wiechert MA, Miner TL, Crosby S, Ciccone R, Willatt L, *et al*: **Clinical and molecular delineation of the 17q21.31 microdeletion syndrome.** *J Med Genet* 2008, **45:**710–720.
46. Goldstein DB: **Common genetic variation and human traits.** *N Engl J Med* 2009, **360:**1696–1698.
47. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA: **Potential etiologic and functional implications of genome-wide association loci for human diseases and traits.** *Proc Natl Acad Sci U S A* 2009, **106:**9362–9367.
48. Maraganore DM, de Andrade M, Lesnick TG, Strain KJ, Farrer MJ, Rocca WA, Pant PV, Frazer KA, Cox DR, Ballinger DG: **High-resolution whole-genome association study of Parkinson disease.** *Am J Hum Genet* 2005, **77:**685–693.
49. Fung HC, Scholz S, Matarin M, Simon-Sanchez J, Hernandez D, Britton A, Gibbs JR, Langefeld C, Stiegert ML, Schymick J, Okun MS, Mandel RJ, Fernandez HH, Foote KD, Rodríguez RL, Peckham E, De Vrieze FW, Gwinn-Hardy K, Hardy JA, Singleton A: **Genome-wide genotyping in Parkinson's disease and neurologically normal controls: first stage analysis and public release of data.** *Lancet Neurol* 2006, **5:**911–916.
50. Pankratz N, Wilk JB, Latourelle JC, DeStefano AL, Halter C, Pugh EW, Doheny KF, Gusella JF, Nichols WC, Foroud T, Myers RH, PSG-PROGENI and GenePD Investigators, Coordinators and Molecular Genetic Laboratories: **Genome-wide association study for susceptibility genes contributing to familial Parkinson disease.** *Hum Genet* 2009, **124:**593–605.
51. Simón-Sánchez J, Schulte C, Bras JM, Sharma M, Gibbs JR, Berg D, Paisan-Ruiz C, Lichtner P, Scholz SW, Hernandez DG, Krüger R, Federoff M, Klein C, Goate A, Perlmutter J, Bonin M, Nalls MA, Illig T, Gieger C, Houlden H, Steffens M, Okun MS, Racette BA, Cookson MR, Foote KD, Fernandez HH, Traynor BJ, Schreiber S, Arepalli S, Zonozi R, *et al*: **Genome-wide association study reveals genetic risk underlying Parkinson's disease.** *Nat Genet* 2009, **41:**1308–1312.
52. Satake W, Nakabayashi Y, Mizuta I, Hirota Y, Ito C, Kubo M, Kawaguchi T, Tsunoda T, Watanabe M, Takeda A, Tomiyama H, Nakashima K, Hasegawa K, Obata F, Yoshikawa T, Kawakami H, Sakoda S, Yamamoto M, Hattori N, Murata M, Nakamura Y, Toda T: **Genome-wide association study identifies common variants at four loci as genetic risk factors for Parkinson's disease.** *Nat Genet* 2009, **41:**1303–1307.
53. Hamza TH, Zabetian CP, Tenesa A, Laederach A, Montimurro J, Yearout D, Kay DM, Doheny KF, Paschall J, Pugh E, Kusel VI, Collura R, Roberts J, Griffith A, Samii A, Scott WK, Nutt J, Factor SA, Payami H: **Common genetic variation in the HLA region is associated with late-onset sporadic Parkinson's disease.** *Nat Genet* 2010, **42:**781–785.
54. Simón-Sánchez J, van Hilten JJ, van de Warrenburg B, Post B, Berendse HW, Arepalli S, Hernandez DG, de Bie RM, Velseboer D, Scheffer H, Bloem B, van Dijk KD, Rivadeneira F, Hofman A, Uitterlinden AG, Rizzu P, Bochdanovits Z, Singleton AB, Heutink P: **Genome-wide association study confirms extant PD risk loci among the Dutch.** *Eur J Hum Genet* 2011, **19:**655–661.
55. International Parkinson Disease Genomics Consortium, Nalls MA, Plagnol V, Hernandez DG, Sharma M, Sheerin UM, Saad M, Simón-Sánchez J, Schulte C, Lesage S, Sveinbjörnsdóttir S, Stefánsson K, Martinez M, Hardy J, Heutink P, Brice A, Gasser T, Singleton AB, Wood NW: **Imputation of sequence variants for identification of genetic risks for Parkinson's disease: a meta-analysis of genome-wide association studies.** *Lancet* 2011, **377:**641–649.
56. Polymeropoulos MH, Lavedan C, Leroy E, Ide SE, Dehejia A, Dutra A, Pike B, Root H, Rubenstein J, Boyer R, Stenroos ES, Chandrasekharappa S, Athanassiadou A, Papapetropoulos T, Johnson WG, Lazzarini AM, Duvoisin RC, Di Iorio G, Golbe LI, Nussbaum RL: **Mutation in the alpha-synuclein gene identified in families with Parkinson's disease.** *Science* 1997, **276:**2045–2047.
57. Farrer M, Maraganore DM, Lockhart P, Singleton A, Lesnick TG, de Andrade M, West A, de Silva R, Hardy J, Hernandez D: **alpha-Synuclein gene haplotypes are associated with Parkinson's disease.** *Hum Mol Genet* 2001, **10:**1847–1851.
58. Singleton AB, Farrer M, Johnson J, Singleton A, Hague S, Kachergus J, Hulihan M, Peuralinna T, Dutra A, Nussbaum R, Lincoln S, Crawley A, Hanson M, Maraganore D, Adler C, Cookson MR, Muenter M, Baptista M, Miller D, Blancato J, Hardy J, Gwinn-Hardy K: **alpha-Synuclein locus triplication causes Parkinson's disease.** *Science* 2003, **302:**841.
59. Zhang CX, Engqvist-Goldstein AE, Carreno S, Owen DJ, Smythe E, Drubin DG: **Multiple roles for cyclin G-associated kinase in clathrin-mediated sorting events.** *Traffic* 2005, **6:**1103–1113.
60. Kawasaki T, Kobayashi T, Ueyama T, Shirai Y, Saito N: **Regulation of clathrin-dependent endocytosis by diacylglycerol kinase delta: importance of kinase activity and binding to AP2alpha.** *Biochem J* 2008, **409:**471–479.
61. Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, McGee S, Do R, Liu X, Jun G, Kang HM, Jordan D, Leal SM, Gabriel S, Rieder MJ, Abecasis G, Altshuler D, Nickerson DA, Boerwinkle E, Sunyaev S, Bustamante CD, Bamshad MJ, Akey JM, Broad GO, Seattle GO, NHLBI Exome Sequencing Projec: **Evolution and functional impact of rare coding variation from deep sequencing of human exomes.** *Science* 2012, **337:**64–69.
62. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M, Bhattacharjee A, Eichler EE, Bamshad M, Nickerson DA, Shendure J: **Targeted capture and massively parallel sequencing of 12 human exomes.** *Nature* 2009, **461:**272–276.
63. Vilariño-Güell C, Wider C, Ross OA, Dachsel JC, Kachergus JM, Lincoln SJ, Soto-Ortolaza AI, Cobb SA, Wilhoite GJ, Bacon JA, Behrouz B, Melrose HL, Hentati E, Puschmann A, Evans DM, Conibear E, Wasserman WW, Aasly JO, Burkhard PR, Djaldetti R, Ghika J, Hentati F, Krygowska-Wajs A, Lynch T, Melamed E, Rajput A, Rajput AH, Solida A, Wu RM, Uitti RJ, *et al*: **VPS35 mutations in Parkinson disease.** *Am J Hum Genet* 2011, **89:**162–167.

64. Zimprich A, Benet-Pagès A, Struhal W, Graf E, Eck SH, Offman MN, Haubenberger D, Spielberger S, Schulte EC, Lichtner P, Rossle SC, Klopp N, Wolf E, Seppi K, Pirker W, Presslauer S, Mollenhauer B, Katzenschlager R, Foki T, Hotzy C, Reinthaler E, Harutyunyan A, Kralovics R, Peters A, Zimprich F, Brücke T, Poewe W, Auff E, Trenkwalder C, Rost B, *et al*: **A mutation in VPS35, encoding a subunit of the retromer complex, causes late-onset Parkinson disease.** *Am J Hum Genet* 2011, **89:**168–175.
65. Lesage S, Condroyer C, Klebe S, Honore A, Tison F, Brefel-Courbon C, Durr A, Brice A, French Parkinson's Disease Genetics Study G: **Identification of VPS35 mutations replicated in French families with Parkinson disease.** *Neurology* 2012, **78:**1449–1450.
66. Ando M, Funayama M, Li Y, Kashihara K, Murakami Y, Ishizu N, Toyoda C, Noguchi K, Hashimoto T, Nakano N, Sasaki R, Kokubo Y, Kuzuhara S, Ogaki K, Yamashita C, Yoshino H, Hatano T, Tomiyama H, Hattori N: **VPS35 mutation in Japanese patients with typical Parkinson's disease.** *Mov Disord* 2012, **27:**1413–1417.
67. Nuytemans K, Bademci G, Inchausti V, Dressen A, Kinnamon DD, Mehta A, Wang L, Zuchner S, Beecham GW, Martin ER, Scott WK, Vance JM: **Whole exome sequencing of rare variants in EIF4G1 and VPS35 in Parkinson disease.** *Neurology* 2013, **80:**982–989.
68. Di Fonzo A, Dekker MC, Montagna P, Baruzzi A, Yonova EH, Correia Guedes L, Szczerbinska A, Zhao T, Dubbel-Hulsman LO, Wouters CH, de Graaff E, Oyen WJ, Simons EJ, Breedveld GJ, Oostra BA, Horstink MW, Bonifati V: **FBXO7 mutations cause autosomal recessive, early-onset parkinsonian-pyramidal syndrome.** *Neurology* 2009, **72:**240–245.
69. Haack TB, Hogarth P, Kruer MC, Gregory A, Wieland T, Schwarzmayr T, Graf E, Sanford L, Meyer E, Kara E, Cuno SM, Harik SI, Dandu VH, Nardocci N, Zorzi G, Dunaway T, Tarnopolsky M, Skinner S, Frucht S, Hanspal E, Schrander-Stumpel C, Héron D, Mignot C, Garavaglia B, Bhatia K, Hardy J, Strom TM, Boddaert N, Houlden HH, Kurian MA, *et al*: **Exome sequencing reveals de novo WDR45 mutations causing a phenotypically distinct, X-linked dominant form of NBIA.** *Am J Hum Genet* 2012, **91:**1144–1149.
70. Edvardson S, Cinnamon Y, Ta-Shma A, Shaag A, Yim YI, Zenvirt S, Jalas C, Lesage S, Brice A, Taraboulos A, Kaestner KH, Greene LE, Elpeleg O: **A deleterious mutation in DNAJC6 encoding the neuronal-specific clathrin-uncoating co-chaperone auxilin, is associated with juvenile parkinsonism.** *PLoS One* 2012, **7:**e36458.
71. Korvatska O, Strand NS, Berndt JD, Strovas T, Chen DH, Leverenz JB, Kiianitsa K, Mata IF, Karakoc E, Greenup JL, Bonkowski E, Chuang J, Moon RT, Eichler EE, Nickerson DA, Zabetian CP, Kraemer BC, Bird TD, Raskind WH: **Altered splicing of ATP6AP2 causes X-linked parkinsonism with spasticity (XPDS).** *Hum Mol Genet* 2013, **22:**3259–3268.
72. Krebs CE, Karkheiran S, Powell JC, Cao M, Makarov V, Darvish H, Di Paolo G, Walker RH, Shahidi GA, Buxbaum JD, De Camilli P, Yue Z, Paisán-Ruiz C: **The Sac1 domain of SYNJ1 identified mutated in a family with early-onset progressive Parkinsonism with generalized seizures.** *Hum Mutat* 2013, **34:**1200–1207.
73. Verkerk AJ, Pieretti M, Sutcliffe JS, Fu YH, Kuhl DP, Pizzuti A, Reiner O, Richards S, Victoria MF, Zhang FP, Eussen BE, van Ommen GB, Blonden LAJ, Riggins GJ, Chastain JL, Kunst CB, Galjaard H, Caskey CT, Nelson DL, Oostra BA, Warren ST: **Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome.** *Cell* 1991, **65:**905–914.
74. Green RC, Berg JS, Grody WW, Kalia SS, Korf BR, Martin CL, McGuire AL, Nussbaum RL, O'Daniel JM, Ormond KE, Rehm HL, Watson MS, Williams MS, Biesecker LG, American College of Medical Genetics and Genomics: **ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing.** *Genet Med* 2013, **15:**565–574.
75. Weischenfeldt J, Symmons O, Spitz F, Korbel JO: **Phenotypic impact of genomic structural variation: insights from and for human disease.** *Nat Rev Genet* 2013, **14:**125–138.
76. Yang Y, Muzny DM, Reid JG, Bainbridge MN, Willis A, Ward PA, Braxton A, Beuten J, Xia F, Niu Z, Hardison M, Person R, Bekheirnia MR, Leduc MS, Kirby A, Pham P, Scull J, Wang M, Ding Y, Plon SE, Lupski JR, Beaudet AL, Gibbs RA, Eng CM: **Clinical whole-exome sequencing for the diagnosis of mendelian disorders.** *N Engl J Med* 2013, **369:**1502–1511.
77. Klein C, Chuang R, Marras C, Lang AE: **The curious case of phenocopies in families with genetic Parkinson's disease.** *Mov Disord* 2011, **26:**1793–1802.
78. Westphal CH, Chandra SS: **Monomeric synucleins generate membrane curvature.** *J Biol Chem* 2013, **288:**1829–1840.
79. Bartels T, Choi JG, Selkoe DJ: **alpha-Synuclein occurs physiologically as a helically folded tetramer that resists aggregation.** *Nature* 2011, **477:**107–110.
80. Lashuel HA, Overk CR, Oueslati A, Masliah E: **The many faces of alpha-synuclein: from structure and toxicity to therapeutic target.** *Nat Rev Neurosci* 2013, **14:**38–48.
81. Nakamura K, Nemani VM, Wallender EK, Kaehlcke K, Ott M, Edwards RH: **Optical reporters for the conformation of alpha-synuclein reveal a specific interaction with mitochondria.** *J Neurosci* 2008, **28:**12305–12317.
82. Boassa D, Berlanga ML, Yang MA, Terada M, Hu J, Bushong EA, Hwang M, Masliah E, George JM, Ellisman MH: **Mapping the subcellular distribution of alpha-synuclein in neurons using genetically encoded probes for correlated light and electron microscopy: implications for Parkinson's disease pathogenesis.** *J Neurosci* 2013, **33:**2605–2615.
83. Plowey ED, Johnson JW, Steer E, Zhu W, Eisenberg DA, Valentino NM, Liu YJ, Chu CT: **Mutant LRRK2 enhances glutamatergic synapse activity and evokes excitotoxic dendrite degeneration.** *Biochim Biophys Acta* 2014. doi: 10.1016/j.bbadis.2014.05.016.
84. Parisiadou L, Yu J, Sgobio C, Xie C, Liu G, Sun L, Gu XL, Lin X, Crowley NA, Lovinger DM, Cai H: **LRRK2 regulates synaptogenesis and dopamine receptor activation through modulation of PKA activity.** *Nat Neurosci* 2014, **17:**367–376.
85. MacLeod D, Dowman J, Hammond R, Leete T, Inoue K, Abeliovich A: **The familial Parkinsonism gene LRRK2 regulates neurite process morphology.** *Neuron* 2006, **52:**587–593.
86. Dachsel JC, Behrouz B, Yue M, Beevers JE, Melrose HL, Farrer MJ: **A comparative study of Lrrk2 function in primary neuronal cultures.** *Parkinsonism Relat Disord* 2010, **16:**650–655.
87. Winner B, Melrose HL, Zhao C, Hinkle KM, Yue M, Kent C, Braithwaite AT, Ogholikhan S, Aigner R, Winkler J, Farrer MJ, Gage FH: **Adult neurogenesis and neurite outgrowth are impaired in LRRK2 G2019S mice.** *Neurobiol Dis* 2011, **41:**706–716.
88. Matta S, Van Kolen K, da Cunha R, van den Bogaart G, Mandemakers W, Miskiewicz K, De Bock PJ, Morais VA, Vilain S, Haddad D, Delbroek L, Swerts J, Chávez-Gutiérrez L, Esposito G, Daneels G, Karran E, Holt M, Gevaert K, Moechars DW, De Strooper B, Verstreken P: **LRRK2 controls an EndoA phosphorylation cycle in synaptic endocytosis.** *Neuron* 2012, **75:**1008–1021.
89. Stafa K, Tsika E, Moser R, Musso A, Glauser L, Jones A, Biskup S, Xiong Y, Bandopadhyay R, Dawson VL, Dawson TM, Moore DJ: **Functional interaction of Parkinson's disease-associated LRRK2 with members of the dynamin GTPase superfamily.** *Hum Mol Genet* 2014, **23:**2055–2077.
90. Sakaguchi-Nakashima A, Meir JY, Jin Y, Matsumoto K, Hisamoto N: **LRK-1, a *C. elegans* PARK8-related kinase, regulates axonal-dendritic polarity of SV proteins.** *Curr Biol* 2007, **17:**592–598.
91. Tong Y, Shen J: **alpha-synuclein and LRRK2: partners in crime.** *Neuron* 2009, **64:**771–773.
92. Dzamko N, Halliday GM: **An emerging role for LRRK2 in the immune system.** *Biochem Soc Trans* 2012, **40:**1134–1139.
93. Vilariño-Güell C, Rajput A, Milnerwood AJ, Shah B, Szu-Tu C, Trinh J, Yu I, Encarnacion M, Munsie LN, Tapia L, Gustavsson EK, Chou P, Tatarnikov I, Evans DM, Pishotta FT, Volta M, Beccano-Kelly D, Thompson C, Lin MK, Sherman HE, Han HJ, Guenther BL, Wasserman WW, Bernard V, Ross CJ, Appel-Cresswell S, Stoessl AJ, Robinson CA, Dickson DW, Ross OA, *et al*: **DNAJC13 mutations in Parkinson disease.** *Hum Mol Genet* 2014, **23:**1794–1801.
94. Helfer E, Harbour ME, Henriot V, Lakisic G, Sousa-Blin C, Volceanov L, Seaman MN, Gautreau A: **Endosomal recruitment of the WASH complex: active sequences and mutations impairing interaction with the retromer.** *Biol Cell* 2013, **105:**191–207.
95. Seaman MN, Gautreau A, Billadeau DD: **Retromer-mediated endosomal protein sorting: all WASHed up!** *Trends Cell Biol* 2013, **23:**522–528.
96. MacLeod DA, Rhinn H, Kuwahara T, Zolin A, Di Paolo G, McCabe BD, Marder KS, Honig LS, Clark LN, Small SA, Abeliovich A: **RAB7L1 interacts with LRRK2 to modify intraneuronal protein sorting and Parkinson's disease risk.** *Neuron* 2013, **77:**425–439.
97. Settembre C, Fraldi A, Medina DL, Ballabio A: **Signals from the lysosome: a control centre for cellular clearance and energy metabolism.** *Nat Rev Mol Cell Biol* 2013, **14:**283–296.
98. Sidransky E, Nalls MA, Aasly JO, Aharon-Peretz J, Annesi G, Barbosa ER, Bar-Shira A, Berg D, Bras J, Brice A, Chen CM, Clark LN, Condroyer C, De Marco EV, Dürr A, Eblan MJ, Fahn S, Farrer MJ, Fung HC, Gan-Or Z, Gasser T, Gershoni-Baruch R, Giladi N, Griffith A, Gurevich T, Januario C, Kropp P, Lang AE, Lee-Chen GJ, Lesage S, *et al*: **Multicenter analysis of glucocerebrosidase mutations in Parkinson's disease.** *N Engl J Med* 2009, **361:**1651–1661.

99. Bras J, Singleton A, Cookson MR, Hardy J: **Emerging pathways in genetic Parkinson's disease: potential role of ceramide metabolism in Lewy body disease.** *FEBS J* 2008, **275:**5767–5773.
100. Mazzulli JR, Xu YH, Sun Y, Knight AL, McLean PJ, Caldwell GA, Sidransky E, Grabowski GA, Krainc D: **Gaucher disease glucocerebrosidase and alpha-synuclein form a bidirectional pathogenic loop in synucleinopathies.** *Cell* 2011, **146:**37–52.
101. Sardi SP, Clarke J, Viel C, Chan M, Tamsett TJ, Treleaven CM, Bu J, Sweet L, Passini MA, Dodge JC, Yu WH, Sidman RL, Cheng SH, Shihabuddin LS: **Augmenting CNS glucocerebrosidase activity as a therapeutic strategy for parkinsonism and other Gaucher-related synucleinopathies.** *Proc Natl Acad Sci U S A* 2013, **110:**3537–3542.
102. Ramirez A, Heimbach A, Grundemann J, Stiller B, Hampshire D, Cid LP, Goebel I, Mubaidin AF, Wriekat AL, Roeper J, Al-Din A, Hillmer AM, Karsak M, Liss B, Woods CG, Behrens MI, Kubisch C: **Hereditary parkinsonism with dementia is caused by mutations in ATP13A2, encoding a lysosomal type 5 P-type ATPase.** *Nat Genet* 2006, **38:**1184–1191.
103. Schultheis PJ, Fleming SM, Clippinger AK, Lewis J, Tsunemi T, Giasson B, Dickson DW, Mazzulli JR, Bardgett ME, Haik KL, Ekhator O, Chava AK, Howard J, Gannon M, Hoffman E, Chen Y, Prasad V, Linn SC, Tamargo RJ, Westbroek W, Sidransky E, Krainc D, Shull GE: **Atp13a2-deficient mice exhibit neuronal ceroid lipofuscinosis, limited alpha-synuclein accumulation and age-dependent sensorimotor deficits.** *Hum Mol Genet* 2013, **22:**2067–2082.
104. Gregory A, Hayflick S: **Neurodegeneration with brain iron accumulation disorders overview.** In *GeneReviews.* Edited by Pagon RA, Adam MP, Bird TD, Dolan CR, Fong CT, Smith RJH, Stephens K. Seattle: University of Washington; 1993-2014:1–36.
105. Langston JW, Ballard P, Tetrud JW, Irwin I: **Chronic Parkinsonism in humans due to a product of meperidine-analog synthesis.** *Science* 1983, **219:**979–980.
106. Keeney PM, Xie J, Capaldi RA, Bennett JP Jr: **Parkinson's disease brain mitochondrial complex I has oxidatively damaged subunits and is functionally impaired and misassembled.** *J Neurosci* 2006, **26:**5256–5264.
107. Spinazzola A, Invernizzi F, Carrara F, Lamantea E, Donati A, Dirocco M, Giordano I, Meznaric-Petrusa M, Baruffini E, Ferrero I, Zeviani M: **Clinical and molecular features of mitochondrial DNA depletion syndromes.** *J Inherit Metab Dis* 2009, **32:**143–158.
108. Dai Y, Kiselak T, Clark J, Clore E, Zheng K, Cheng A, Kujoth GC, Prolla TA, Maratos-Flier E, Simon DK: **Behavioral and metabolic characterization of heterozygous and homozygous POLG mutator mice.** *Mitochondrion* 2013, **13:**282–291.
109. Kitada T, Asakawa S, Hattori N, Matsumine H, Yamamura Y, Minoshima S, Yokochi M, Mizuno Y, Shimizu N: **Mutations in the parkin gene cause autosomal recessive juvenile parkinsonism.** *Nature* 1998, **392:**605–608.
110. Ko HS, von Coelln R, Sriram SR, Kim SW, Chung KK, Pletnikova O, Troncoso J, Johnson B, Saffary R, Goh EL, Song H, Park BJ, Kim MJ, Kim S, Dawson VL, Dawson TM: **Accumulation of the authentic parkin substrate aminoacyl-tRNA synthetase cofactor, p38/JTV-1, leads to catecholaminergic cell death.** *J Neurosci* 2005, **25:**7968–7978.
111. Fallon L, Belanger CM, Corera AT, Kontogiannea M, Regan-Klapisz E, Moreau F, Voortman J, Haber M, Rouleau G, Thorarinsdottir T, Brice A, van Bergen En Henegouwen PM, Fon EA: **A regulated interaction with the UIM protein Eps15 implicates parkin in EGF receptor trafficking and PI(3)K-Akt signalling.** *Nat Cell Biol* 2006, **8:**834–842.
112. Staropoli JF, McDermott C, Martinat C, Schulman B, Demireva E, Abeliovich A: **Parkin is a component of an SCF-like ubiquitin ligase complex and protects postmitotic neurons from kainate excitotoxicity.** *Neuron* 2003, **37:**735–749.
113. Burchell VS, Nelson DE, Sanchez-Martinez A, Delgado-Camprubi M, Ivatt RM, Pogson JH, Randle SJ, Wray S, Lewis PA, Houlden H, Abramov AY, Hardy J, Wood NW, Whitworth AJ, Laman H, Plun-Favreau H: **The Parkinson's disease-linked proteins Fbxo7 and Parkin interact to mediate mitophagy.** *Nat Neurosci* 2013, **16:**1257–1265.
114. Park J, Lee SB, Lee S, Kim Y, Song S, Kim S, Bae E, Kim J, Shong M, Kim JM, Chung J: **Mitochondrial dysfunction in *Drosophila* PINK1 mutants is complemented by parkin.** *Nature* 2006, **441:**1157–1161.
115. Clark IE, Dodson MW, Jiang C, Cao JH, Huh JR, Seol JH, Yoo SJ, Hay BA, Guo M: ***Drosophila* pink1 is required for mitochondrial function and interacts genetically with parkin.** *Nature* 2006, **441:**1162–1166.
116. Trempe JF, Sauve V, Grenier K, Seirafi M, Tang MY, Menade M, Al-Abdul-Wahid S, Krett J, Wong K, Kozlov G, Nagar B, Fon EA, Gehring K: **Structure of parkin reveals mechanisms for ubiquitin ligase activation.** *Science* 2013, **340:**1451–1455.
117. Koyano F, Okatsu K, Kosako H, Tamura Y, Go E, Kimura M, Kimura Y, Tsuchiya H, Yoshihara H, Hirokawa T, Endo T, Fon EA, Trempe JF, Saeki Y, Tanaka K, Matsuda N: **Ubiquitin is phosphorylated by PINK1 to activate parkin.** *Nature* 2014, **510:**162–166.
118. Hasson SA, Kane LA, Yamano K, Huang CH, Sliter DA, Buehler E, Wang C, Heman-Ackah SM, Hessa T, Guha R, Martin SE, Youle RJ: **High-content genome-wide RNAi screens identify regulators of parkin upstream of mitophagy.** *Nature* 2013, **504:**291–295.
119. McCoy MK, Kaganovich A, Rudenko IN, Ding J, Cookson MR: **Hexokinase activity is required for recruitment of parkin to depolarized mitochondria.** *Hum Mol Genet* 2014, **23:**145–156.
120. Hajek P, Chomyn A, Attardi G: **Identification of a novel mitochondrial complex containing mitofusin 2 and stomatin-like protein 2.** *J Biol Chem* 2007, **282:**5670–5681.
121. Poole AC, Thomas RE, Yu S, Vincow ES, Pallanck L: **The mitochondrial fusion-promoting factor mitofusin is a substrate of the PINK1/parkin pathway.** *PLoS One* 2010, **5:**e10054.
122. Rakovic A, Grunewald A, Voges L, Hofmann S, Orolicki S, Lohmann K, Klein C: **PINK1-interacting proteins: proteomic analysis of overexpressed PINK1.** *Parkinsons Dis* 2011, **2011:**153979.
123. Bonifati V, Rizzu P, van Baren MJ, Schaap O, Breedveld GJ, Krieger E, Dekker MC, Squitieri F, Ibanez P, Joosse M, van Dongen JW, Vanacore N, van Swieten JC, Brice A, Meco G, van Duijn CM, Oostra BA, Heutink P: **Mutations in the DJ-1 gene associated with autosomal recessive early-onset parkinsonism.** *Science* 2003, **299:**256–259.
124. Joselin AP, Hewitt SJ, Callaghan SM, Kim RH, Chung YH, Mak TW, Shen J, Slack RS, Park DS: **ROS-dependent regulation of Parkin and DJ-1 localization during oxidative stress in neurons.** *Hum Mol Genet* 2012, **21:**4888–4903.
125. Gispert S, Ricciardi F, Kurz A, Azizov M, Hoepken HH, Becker D, Voos W, Leuner K, Müller WE, Kudin AP, Kunz WS, Zimmermann A, Roeper J, Wenzel D, Jendrach M, García-Arencíbia M, Fernández-Ruiz J, Huber L, Rohrer H, Barrera M, Reichert AS, Rüb U, Chen A, Nussbaum RL, Auburger G: **Parkinson phenotype in aged PINK1-deficient mice is accompanied by progressive mitochondrial dysfunction in absence of neurodegeneration.** *PLoS One* 2009, **4:**e5777.
126. Goldberg MS, Pisani A, Haburcak M, Vortherms TA, Kitada T, Costa C, Tong Y, Martella G, Tscherter A, Martins A, Bernardi G, Roth BL, Pothos EN, Calabresi P, Shen J: **Nigrostriatal dopaminergic deficits and hypokinesia caused by inactivation of the familial Parkinsonism-linked gene DJ-1.** *Neuron* 2005, **45:**489–496.
127. Goldberg MS, Fleming SM, Palacino JJ, Cepeda C, Lam HA, Bhatnagar A, Meloni EG, Wu N, Ackerson LC, Klapstein GJ, Gajendiran M, Roth BL, Chesselet MF, Maidment NT, Levine MS, Shen J: **Parkin-deficient mice exhibit nigrostriatal deficits but not loss of dopaminergic neurons.** *J Biol Chem* 2003, **278:**43628–43635.
128. Shin JH, Ko HS, Kang H, Lee Y, Lee YI, Pletinkova O, Troconso JC, Dawson VL, Dawson TM: **PARIS (ZNF746) repression of PGC-1α contributes to neurodegeneration in Parkinson's disease.** *Cell* 2011, **144:**689–702.
129. Martin LJ, Pan Y, Price AC, Sterling W, Copeland NG, Jenkins NA, Price DL, Lee MK: **Parkinson's disease alpha-synuclein transgenic mice develop neuronal mitochondrial degeneration and cell death.** *J Neurosci* 2006, **26:**41–50.
130. Cooper O, Seo H, Andrabi S, Guardia-Laguarta C, Graziotto J, Sundberg M, McLean JR, Carrillo-Reid L, Xie Z, Osborn T, Hargus G, Deleidi M, Lawson T, Bogetofte H, Perez-Torres E, Clark L, Moskowitz C, Mazzulli J, Chen L, Volpicelli-Daley L, Romero N, Jiang H, Uitti RJ, Huang Z, Opala G, Scarffe LA, Dawson VL, Klein C, Feng J, Ross OA, *et al*: **Pharmacological rescue of mitochondrial deficits in iPSC-derived neural cells from patients with familial Parkinson's disease.** *Sci Transl Med* 2012, **4:**141ra190.
131. Ramonet D, Daher JP, Lin BM, Stafa K, Kim J, Banerjee R, Westerlund M, Pletnikova O, Glauser L, Yang L, Liu Y, Swing DA, Beal MF, Troncoso JC, McCaffery JM, Jenkins NA, Copeland NG, Galter D, Thomas B, Lee MK, Dawson TM, Dawson VL, Moore DJ: **Dopaminergic neuronal loss, reduced neurite complexity and autophagic abnormalities in transgenic mice expressing G2019S mutant LRRK2.** *PLoS One* 2011, **6:**e18568.
132. Wang X, Yan MH, Fujioka H, Liu J, Wilson-Delfosse A, Chen SG, Perry G, Casadesus G, Zhu X: **LRRK2 regulates mitochondrial dynamics and function through direct interaction with DLP1.** *Hum Mol Genet* 2012, **21:**1931–1944.

133. Zhi D, Chen R: **Statistical guidance for experimental design and data analysis of mutation detection in rare monogenic mendelian diseases by exome sequencing.** *PLoS One* 2012, **7:**e31358.
134. Choy RW, Park M, Temkin P, Herring BE, Marley A, Nicoll RA, von Zastrow M: **Retromer mediates a discrete route of local membrane delivery to dendrites.** *Neuron* 2014, **82:**55–62.
135. Stocchi F, Olanow CW: **Obstacles to the development of a neuroprotective therapy for Parkinson's disease.** *Mov Disord* 2013, **28:**3–7.
136. Spillantini MG, Schmidt ML, Lee VM, Trojanowski JQ, Jakes R, Goedert M: **Alpha-synuclein in Lewy bodies.** *Nature* 1997, **388:**839–840.
137. Abbas N, Lücking CB, Ricard S, Dürr A, Bonifati V, De Michele G, Bouley S, Vaughan JR, Gasser T, Marconi R, Broussolle E, Brefel-Courbon C, Harhangi BS, Oostra BA, Fabrizio E, Böhme GA, Pradier L, Wood NW, Filla A, Meco G, Denefle P, Agid Y, Brice A: **A wide variety of mutations in the parkin gene are responsible for autosomal recessive parkinsonism in Europe. French Parkinson's Disease Genetics Study Group and the European Consortium on Genetic Susceptibility in Parkinson's Disease.** *Hum Mol Genet* 1999, **8:**567–574.
138. van de Warrenburg BP, Lammens M, Lucking CB, Denefle P, Wesseling P, Booij J, Praamstra P, Quinn N, Brice A, Horstink MW: **Clinical and pathologic abnormalities in a family with parkinsonism and parkin gene mutations.** *Neurology* 2001, **56:**555–557.
139. Farrer M, Chan P, Chen R, Tan L, Lincoln S, Hernandez D, Forno L, Gwinn-Hardy K, Petrucelli L, Hussey J, Singleton A, Tanner C, Hardy J, Langston JW: **Lewy bodies and parkinsonism in families with parkin mutations.** *Ann Neurol* 2001, **50:**293–300.
140. Krüger R, Vieira-Saecker AM, Kuhn W, Berg D, Müller T, Kühnl N, Fuchs GA, Storch A, Hungs M, Woitalla D, Przuntek H, Epplen JT, Schöls L, Riess O: **Increased susceptibility to sporadic Parkinson's disease by a certain combined alpha-synuclein/apolipoprotein E genotype.** *Ann Neurol* 1999, **45:**611–617.
141. Maraganore DM, Hernandez DG, Singleton AB, Farrer MJ, McDonnell SK, Hutton ML, Hardy JA, Rocca WA: **Case-control study of the extended tau gene haplotype in Parkinson's disease.** *Ann Neurol* 2001, **50:**658–661.
142. Höglinger GU, Melhem NM, Dickson DW, Sleiman PM, Wang LS, Klei L, Rademakers R, de Silva R, Litvan I, Riley DE, van Swieten JC, Heutink P, Wszolek ZK, Uitti RJ, Vandrovcova J, Hurtig HI, Gross RG, Maetzler W, Goldwurm S, Tolosa E, Borroni B, Pastor P, PSP Genetics Study Group, Cantwell LB, Han MR, Dillman A, van der Brug MP, Gibbs JR, Cookson MR, Hernandez DG, *et al*: **Identification of common variants influencing risk of the tauopathy progressive supranuclear palsy.** *Nat Genet* 2011, **43:**699–705.
143. Farrer M, Kachergus J, Forno L, Lincoln S, Wang DS, Hulihan M, Maraganore D, Gwinn-Hardy K, Wszolek Z, Dickson D, Langston JW: **Comparison of kindreds with parkinsonism and alpha-synuclein genomic multiplications.** *Ann Neurol* 2004, **55:**174–179.
144. Chartier-Harlin MC, Kachergus J, Roumier C, Mouroux V, Douay X, Lincoln S, Levecque C, Larvor L, Andrieux J, Hulihan M, Waucquier N, Defebvre L, Amouyel P, Farrer M, Destée A: **alpha-synuclein locus duplication as a cause of familial Parkinson's disease.** *Lancet* 2004, **364:**1167–1169.
145. Ibanez P, Bonnet AM, Debarges B, Lohmann E, Tison F, Pollak P, Agid Y, Durr A, Brice A: **Causal relation between alpha-synuclein gene duplication and familial Parkinson's disease.** *Lancet* 2004, **364:**1169–1171.
146. Nishioka K, Ross OA, Ishii K, Kachergus JM, Ishiwata K, Kitagawa M, Kono S, Obi T, Mizoguchi K, Inoue Y, Imai H, Takanashi M, Mizuno Y, Farrer MJ, Hattori N: **Expanding the clinical phenotype of SNCA duplication carriers.** *Mov Disord* 2009, **24:**1811–1819.
147. Valente EM, Abou-Sleiman PM, Caputo V, Muqit MM, Harvey K, Gispert S, Ali Z, Del Turco D, Bentivoglio AR, Healy DG, Albanese A, Nussbaum R, González-Maldonado R, Deller T, Salvi S, Cortelli P, Gilks WP, Latchman DS, Harvey RJ, Dallapiccola B, Auburger G, Wood NW: **Hereditary early-onset Parkinson's disease caused by mutations in PINK1.** *Science* 2004, **304:**1158–1160.
148. Kachergus J, Mata IF, Hulihan M, Taylor JP, Lincoln S, Aasly J, Gibson JM, Ross OA, Lynch T, Wiley J, Payami H, Nutt J, Maraganore DM, Czyzewski K, Styczynska M, Wszolek ZK, Farrer MJ, Toft M: **Identification of a novel LRRK2 mutation linked to autosomal dominant parkinsonism: evidence of a common founder across European populations.** *Am J Hum Genet* 2005, **76:**672–680.
149. Chartier-Harlin MC, Dachsel JC, Vilariño-Güell C, Lincoln SJ, Leprêtre F, Hulihan MM, Kachergus J, Milnerwood AJ, Tapia L, Song MS, Le Rhun E, Mutez E, Larvor L, Duflot A, Vanbesien-Mailliot C, Kreisler A, Ross OA, Nishioka K, Soto-Ortolaza AI, Cobb SA, Melrose HL, Behrouz B, Keeling BH, Bacon JA, Hentati E, Williams L, Yanagiya A, Sonenberg N, Lockhart PJ, Zubair AC, *et al*: **Translation initiator EIF4G1 mutations in familial Parkinson disease.** *Am J Hum Genet* 2011, **89:**398–406.
150. Lill CM, Roehr JT, McQueen MB, Kavvoura FK, Bagade S, Schjeide BM, Schjeide LM, Meissner E, Zauft U, Allen NC, Liu T, Schilling M, Anderson KJ, Beecham G, Berg D, Biernacka JM, Brice A, DeStefano AL, Do CB, Eriksson N, Factor SA, Farrer MJ, Foroud T, Gasser T, Hamza T, Hardy JA, Heutink P, Hill-Burns EM, Klein C, Latourelle JC, *et al*: **Comprehensive research synopsis and systematic meta-analyses in Parkinson's disease genetics: The PDGene database.** *PLoS Genet* 2012, **8:**e1002548.
151. Hatano Y, Li Y, Sato K, Asakawa S, Yamamura Y, Tomiyama H, Yoshino H, Asahina M, Kobayashi S, Hassin-Baer S, Lu CS, Ng AR, Rosales RL, Shimizu N, Toda T, Mizuno Y, Hattori N: **Novel PINK1 mutations in early-onset parkinsonism.** *Ann Neurol* 2004, **56:**424–427.
152. Paisan-Ruiz C, Bhatia KP, Li A, Hernandez D, Davis M, Wood NW, Hardy J, Houlden H, Singleton A, Schneider SA: **Characterization of PLA2G6 as a locus for dystonia-parkinsonism.** *Ann Neurol* 2009, **65:**19–23.
153. Lee HJ, Suk JE, Bae EJ, Lee JH, Paik SR, Lee SJ: **Assembly-dependent endocytosis and clearance of extracellular alpha-synuclein.** *Int J Biochem Cell Biol* 2008, **40:**1835–1849.

doi:10.1186/gm566
Cite this article as: Lin and Farrer: Genetics and genomics of Parkinson's disease. *Genome Medicine* 2014 **6**:48.

Hennart *et al. Genome Medicine* (2020) 12:107
https://doi.org/10.1186/s13073-020-00805-7

Genome Medicine

RESEARCH **Open Access**

Population genomics and antimicrobial resistance in *Corynebacterium diphtheriae*

Melanie Hennart[1,2], Leonardo G. Panunzi[1,3], Carla Rodrigues[1], Quentin Gaday[4], Sarah L. Baines[5], Marina Barros-Pinkelnig[1], Annick Carmi-Leroy[1,6], Melody Dazas[1,6], Anne Marie Wehenkel[4], Xavier Didelot[7], Julie Toubiana[1,6,8], Edgar Badell[1,6] and Sylvain Brisse[1,6*]

Abstract

Background: *Corynebacterium diphtheriae*, the agent of diphtheria, is a genetically diverse bacterial species. Although antimicrobial resistance has emerged against several drugs including first-line penicillin, the genomic determinants and population dynamics of resistance are largely unknown for this neglected human pathogen.

Methods: Here, we analyzed the associations of antimicrobial susceptibility phenotypes, diphtheria toxin production, and genomic features in *C. diphtheriae*. We used 247 strains collected over several decades in multiple world regions, including the 163 clinical isolates collected prospectively from 2008 to 2017 in France mainland and overseas territories.

Results: Phylogenetic analysis revealed multiple deep-branching sublineages, grouped into a Mitis lineage strongly associated with diphtheria toxin production and a largely toxin gene-negative Gravis lineage with few toxin-producing isolates including the 1990s ex-Soviet Union outbreak strain. The distribution of susceptibility phenotypes allowed proposing ecological cutoffs for most of the 19 agents tested, thereby defining acquired antimicrobial resistance. Penicillin resistance was found in 17.2% of prospective isolates. Seventeen (10.4%) prospective isolates were multidrug-resistant (≥ 3 antimicrobial categories), including four isolates resistant to penicillin and macrolides. Homologous recombination was frequent (r/m = 5), and horizontal gene transfer contributed to the emergence of antimicrobial resistance in multiple sublineages. Genome-wide association mapping uncovered genetic factors of resistance, including an accessory penicillin-binding protein (PBP2m) located in diverse genomic contexts. Gene *pbp2m* is widespread in other *Corynebacterium* species, and its expression in *C. glutamicum* demonstrated its effect against several beta-lactams. A novel 73-kb *C. diphtheriae* multiresistance plasmid was discovered.

Conclusions: This work uncovers the dynamics of antimicrobial resistance in *C. diphtheriae* in the context of phylogenetic structure, biovar, and diphtheria toxin production and provides a blueprint to analyze re-emerging diphtheria.

Keywords: *C. diphtheriae*, Antibiotic resistance, Genome-wide association study, Phylogeny, Mobile genetic element, Biovar

---

* Correspondence: sbrisse@pasteur.fr
[1]Institut Pasteur, Biodiversity and Epidemiology of Bacterial Pathogens, Paris, France
[6]Institut Pasteur, National Reference Center for Corynebacteria of the Diphtheriae Complex, Paris, France
Full list of author information is available at the end of the article

© The Author(s). 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Background

Diphtheria, if untreated, is one of the most severe bacterial infections of humans. It typically affects the upper respiratory tract causing pseudomembrane formation, sometimes leading to suffocation and death. The infection can be complicated by systemic symptoms, caused by the diphtheria toxin. Other forms of disease are skin and invasive infections, including endocarditis [1, 2].

The agent of diphtheria is *Corynebacterium diphtheriae*, a member of the phylum Actinomycetes [3, 4]. The diphtheria toxin, encoded by the *tox* gene, is carried by lysogenized corynephages within the chromosome of some *C. diphtheriae* strains [5, 6]. Concern exists about the possibility of lysogenic conversion of previously non-toxigenic strains during colonization, infection, or transmission chains [7]. However, knowledge on the microevolutionary dynamics between *tox*-positive and *tox*-negative strains is limited. The high genetic diversity of *C. diphtheriae* strains underlies their variable colonization, adhesion, and pathogenicity properties [8–10]. Although three main biovars (Mitis, Gravis, and Belfanti) are distinguished since the 1950s, their phylogenetic relationships are poorly defined [11–13].

Diphtheria used to be one of the deadliest infections in young children but has been largely controlled by vaccination with the highly effective toxoid vaccine [14]. Even so, thousands of cases of diphtheria are still reported annually [15], and large outbreaks can quickly follow the disruption of public health systems [14, 16–18]. In countries with high vaccination coverage, diphtheria cases are associated with travel and migration from endemic regions [19–21]. As diphtheria vaccination is performed using an inactivated form of diphtheria toxin, it is not considered to prevent asymptomatic colonization and silent transmission of the pathogen, which still circulates and is the object of intense epidemiological surveillance [4]. However, vaccine preparations may include other antigens, and the impact of vaccination on *C. diphtheriae* evolution deserves further studies [22].

Clinical management of infections with toxigenic isolates includes treatment with diphtheria antitoxin (DAT), which can prevent or reduce the systemic effects of the toxin [4]. Nevertheless, antimicrobial treatment is critical in the clinical management of both *tox*-positive and *tox*-negative infections, as it contributes to the elimination of the bacteria within the patient and limits transmission to novel individuals [23]. With DAT production being threatened [24], antimicrobial treatment might become even more critical in diphtheria therapy.

Penicillin is the first-line therapeutics to treat diphtheria, with erythromycin being recommended in case of allergy [25]. Both antimicrobial agents are generally effective for the treatment of diphtheria [23, 26]. However, reduced susceptibility or full resistance of *C. diphtheriae* to penicillin, erythromycin, and other antimicrobial agents, sometimes combined in multidrug-resistant strains, has been reported from multiple world regions [23, 26–35].

Antimicrobial resistance genes have been described in *C. diphtheriae*, including the erythromycin resistance gene *ermX* on plasmid pNG2 [36] and genes *dfrA16*, *qacH*, and *sul1* carried on a class 1 integron, mobilized by IS*6100* [37]. However, the prevalence and phylogenetic distribution of resistance genes in *C. diphtheriae* clinical isolates are unknown. Six chromosomal penicillin-binding proteins (PBPs) have been reported in *C. diphtheriae* [38], but so far, no association between *pbp* or other genetic variation and penicillin resistance has been described. Understanding the genetic basis of antimicrobial resistance in *C. diphtheriae* would improve our ability to diagnose and track its spread.

The aims of this study were (i) to characterize antimicrobial resistance phenotypes in a large collection of *C. diphtheriae* strains with diverse geographical and temporal origins and to uncover genomic determinants of resistance and (ii) to analyze the population structure of *C. diphtheriae* and define the associations between antimicrobial resistance, diphtheria toxin production, biovars, and phylogenetic sublineages.

Methods

Corynebacterium diphtheriae isolates and strains

A collection of 247 *C. diphtheriae* isolates were included (Additional file 1: Table S1), corresponding to three subsets. First, we included 163 clinical isolates (recent clinical isolates subset, Additional file 1: Table S1) collected prospectively between 2008 and 2017 by the French National Reference Center for Corynebacteria of the *Corynebacterium diphtheriae* complex (NRC-CCD). These isolates represented all isolates received at the NRC-CCD that corresponded to the *C. diphtheriae* species (*C. belfantii*, *C. ulcerans*, *C. pseudotuberculosis*, or other corynebacteria were excluded). They were collected from cutaneous (*n* = 136), respiratory (*n* = 23), and other (bones, blood; *n* = 4) infections. Of these, 74 were from Mainland France and 89 from French overseas territories, including Mayotte (*n* = 50), New Caledonia (*n* = 19), La Réunion Island (*n* = 11), French Guiana (*n* = 4), French Polynesia (*n* = 3), and Guadeloupe (*n* = 1); one isolate received from Institut Pasteur in Madagascar was also included (Fig. 1). Four isolates from New Caledonia collected between 2002 and 2006 (02-0322, 02-0338, 03-1641, and 06-1569) were included in a previous study [39]; the trimethoprim- and sulfamethoxazole-resistant isolate FRC0024 was previously shown to harbor an integron with gene *drfA16* [37].

Second, we included 15 clinical isolates collected in France between 1981 and 1991, 11 of which had been deposited in the Collection de l'Institut Pasteur (CIP;

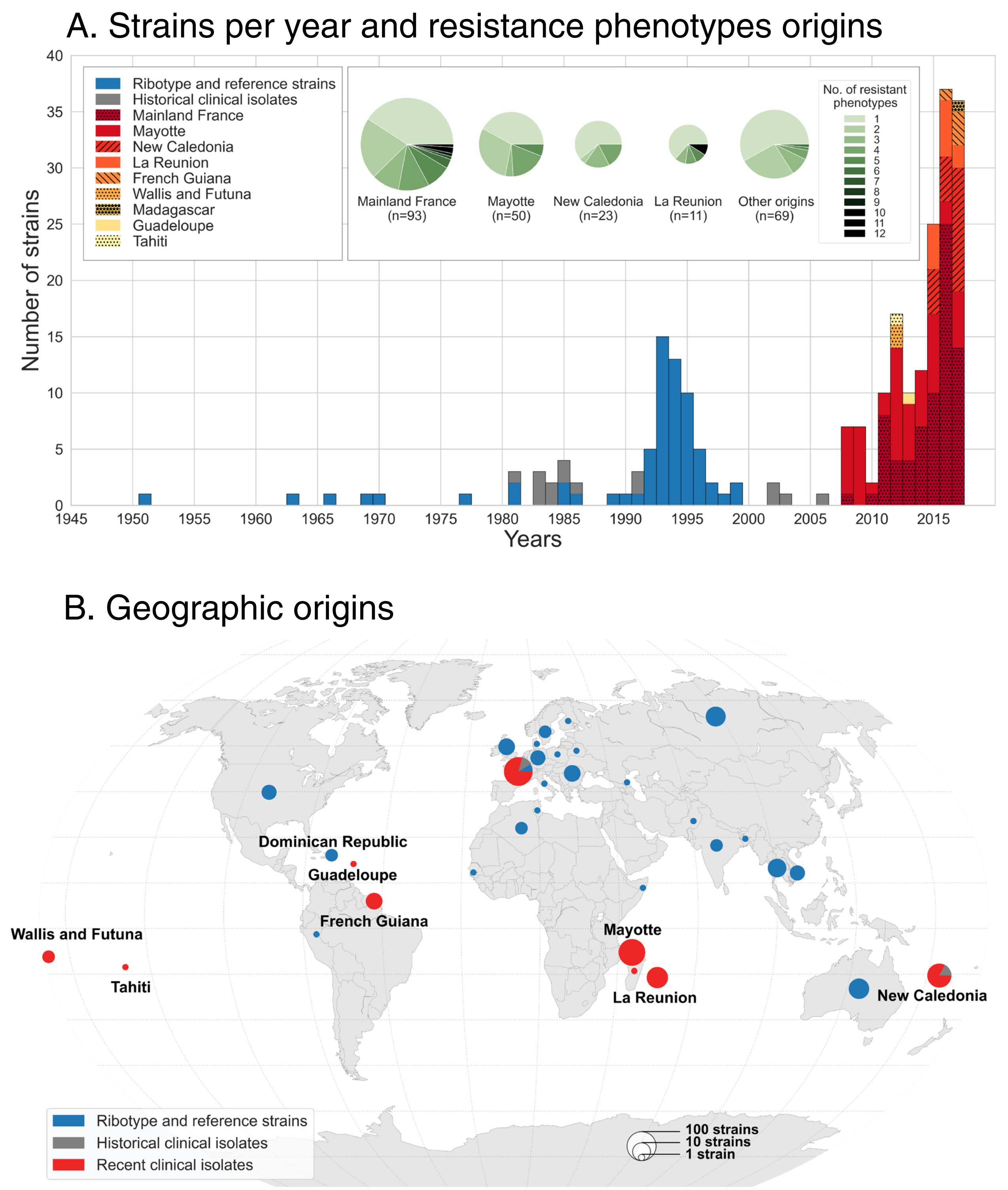


Fig. 1 Temporal and geographical distribution of strains studied. **a** The number of strains per year, 1945–2017. The 2008–2017 clinical isolates are represented in red- or yellow-shaded patterns (see key), whereas the older clinical isolates are in gray and the reference strains in blue. The inset shows the pie charts with the frequency of resistance phenotypes among strains from the four most represented geographic origins; the remaining ones are pooled in the right-most pie chart. **b** Geographic origins of strains from the three subsets

historical clinical isolates subset in Additional file 1: Table S1).

Third, to increase the genetic diversity and geographic range of the sample, the 65 available reference strains of ribotypes that belong to *C. diphtheriae* were included [40]. These reference strains represent an international collection of isolates collected over several decades and originating from multiple world regions including the Americas, Europe, Asia, Africa, and Oceania. Our subcultures of these strains were controlled for *tox* gene presence, toxin production, and biovar, leading to modifications of published characteristics in some instances (Additional file 1: Table S1; Fig. 1). Finally, four reference strains were included: strain NCTC13129, which is used as genomic sequence reference [38]; strain NCTC10648, which is used as the *tox*-positive and toxinogenic reference strain in PCR and Elek tests, respectively; strain NCTC11397^{T}, which is the taxonomic type strain of the *C. diphtheriae* species; and the vaccine production strain PW8, which corresponds to CIP A102 [41]. This third subset is referred to as "ribotype and reference strains" subset (Additional file 1: Table S1).

Bacterial cultures, identification, and biovar

Bacteria were cultivated on Trypto-Casein-Soy (TCS) agar for 24 h at 35–37 °C. Bacterial identification was performed at the NRC-CCD as described previously [42] by multiplex polymerase chain reaction (PCR) combining a *dtxR* gene fragment specific for *C. diphtheriae* and a multiplex PCR that targets a fragment of the *pld* gene specific for *C. pseudotuberculosis*, the gene *rpoB* (amplified in all species of the *C. diphtheriae* complex) and a fragment of 16S rRNA gene specific for *C. pseudotuberculosis* and *C. ulcerans*, respectively. Isolates collected since 2014 were confirmed as *C. diphtheriae* by matrix-assisted laser desorption-ionization time-of-flight mass spectrometry (MALDI-TOF MS) using Bruker technology. In order to exclude strains initially identified as *C. diphtheriae* but now classified as *C. belfantii* [42] or *C. rouxii* [43], genome-wide average nucleotide identity (ANI) was used as described previously [42]. Strains were characterized biochemically for pyrazinamidase, urease, and nitrate reductase and for utilization of maltose and trehalose using API Coryne strips (BioMérieux, Marcy l'Etoile, France) and the Rosco Diagnostica reagents (Eurobio, Les Ulis, France). The Hiss serum water test was used for glycogen fermentation. The biovar of isolates was determined based on the combination of nitrate reductase (positive in Mitis and Gravis, negative in Belfanti) and glycogen fermentation (positive in Gravis only). The rare biovar Intermedius was not identified, as its distinction from other biovars is based on colony morphology, which is considered subjective, or on lipophily, which was not tested.

Determination of the presence of the *tox* gene

Determination of the diphtheria toxin gene (*tox* gene) presence was achieved by a conventional *tox* PCR assay [44], while its phenotypic production was assessed by the modified Elek test [45]. We also confirmed *tox* PCR results by BLASTN (query: *tox* gene sequence from strain NCTC13129, RefSeq accession number: DIP_RS12515) analysis of the genomic assemblies.

Antimicrobial susceptibility testing

Phenotypic susceptibility was tested for the following agents: penicillin G (10 IU), amoxicillin, oxacillin, cefotaxime, imipenem, erythromycin, azithromycin, clarithromycin, spiramycin, pristinamycin, kanamycin, gentamicin, rifampicin, tetracycline, ciprofloxacin, clindamycin, sulfonamide, trimethoprim, and trimethoprim + sulfamethoxazole. The 19 antimicrobial agents tested (Table S1) corresponded to seven classes, as described hereafter: *β*-lactams: penicillin G (PEN), amoxicillin (AMX), oxacillin (OXA), cefotaxime (CFT), imipenem (IMP); macrolides: azithromycin (AZM), clarithromycin (CLR), erythromycin (ERT), and spiramycin (SPR); lincosamides: clindamycin (CLD); streptogramins: pristinamycin (PRT); aminoglycosides: gentamicin (GEN) and kanamycin (KAN); folate pathway inhibitors: sulfonamide (SUL), trimethoprim (TMP), and trimethoprim + sulfamethoxazole (cotrimoxazole, TMP-STX); ansamycins: rifampicin (RIF); tetracyclines: tetracycline (TET); and fluoroquinolones: ciprofloxacin (CIP).

Antimicrobial susceptibility was determined using the disk diffusion method with impregnated paper disks (Bio-Rad, Marnes-la-Coquette, France) on Mueller Hinton agar plates supplemented with 5% of sheep blood and 20 mg/L β-NAD, as recommended. Minimum inhibitory concentrations (MICs) were determined using E-test strips (BioMerieux, Marcy l'Etoile, France). The control strain used is *Streptococcus pneumoniae* ATCC 49619. The zone diameter (ZD) data were interpreted into S, I, and R categories in the following way. First, we used the CA-SFM/EUCAST V.1.0 (January 2019) document (https://www.sfm-microbiologie.org/wp-content/uploads/2019/02/CASFM2019_V1.0.pdf), which contains interpretative criteria for *Corynebacterium* spp. only for CIP, GEN, CLD, TET, RIF, and TMP-STX. Second, for the other agents, we used the interpretative criteria published in Table III of the CA-SFM 2013 recommendations (https://resapath.anses.fr/resapath_uploadfiles/files/Documents/2013_CASFM.pdf). Note that for RIF, we used the 2013 breakpoints, as they fitted better with the observed distribution of ZD values. Clarithromycin breakpoints were taken from those for erythromycin, as recommended. The breakpoint for oxacillin was derived from the one used for *Staphylococcus* spp. ZD interpretation breakpoints are given in Additional file 1: Table S2.

Penicillin susceptibility was initially determined using 10-UI (6 μg) disks (resistance breakpoint, 18 mm), but CA-SFM/EUCAST recommendations were changed in 2014 to use 1-UI disks, while the resistance breakpoint was increased from 18 to 29 mm. As all *C. diphtheriae* strains end up in the resistant category following this recommendation, E-test strips were used to define the penicillin MIC since 2014 (Additional file 1: Table S2: isolates starting from FRC0259); the EUCAST breakpoint of 0.125 g/L was used as a cutoff. Penicillin E-test was also performed systematically for strains tested as resistant before 2014 (using 10-UI disks), as well as for some susceptible isolates (Additional file 1: Table S1).

Multidrug-resistant *C. diphtheriae* (MDR-DIP) were defined as strains resistant to three or more antimicrobial agent categories (defined in Additional file 1: Table S2), excluding intrinsic resistance to fosfomycin [46]. Note that we used ecological cutoffs rather than currently proposed clinical breakpoints (see Additional file 1: Table S2 for a comparison of both types of breakpoints).

Whole-genome sequencing by Illumina and Oxford Nanopore Technologies

DNA was extracted from broth cultures, by making use of DNeasy Blood & Tissue Kit (QIAGEN, Hilden, Germany). However, a lysis step was added to the extraction protocol described by the manufacturer as previously described [47]: a 1-μL loopful of bacterial colonies was emulsified in 180 μL of lysis buffer containing 20 mM Tris-HCl, pH 8, 2 mM EDTA, 1.2% Triton X-100, and 20 mg/mL lysozyme, in a DNase/RNase-free 1.5-mL Eppendorf tube and incubated in a heating block at 37 °C for 1 h, with mixing every 20 min. After extraction, DNA concentration was measured with the Qubit 3.0 Fluorometer (Invitrogen), employing the Qubit dsDNA BR Assay Kit (Invitrogen). Besides, the DNA quality was verified using a D-One spectrophotometer (Nanodrop). Multiplexed paired-end libraries (2 × 150 bp) were prepared using the Nextera XT DNA kit (Illumina, San Diego, CA, USA) and eventually sequenced with an Illumina NextSeq-500 instrument at a minimum of 50× coverage depth. Trimming and clipping were performed using AlienTrimmer v0.4.0 [48]. Redundant or over-represented reads were reduced using the khmer software package v1.3 [49]. Finally, sequencing errors were corrected using Musket v1.1 [50]. A de novo assembly was performed for each strain using SPAdes v3.12.0 [51]. The genomic sequences of the four reference strains were retrieved from public repositories (Additional file 1: Table S1).

Additionally, the multidrug-resistant isolate FRC0402 was subjected to long-read sequencing using Oxford Nanopore Technologies (ONT). Genomic DNA was extracted using the phenol-chloroform protocol combined with Phase Lock Gel tubes (Qiagen GmbH). Libraries were prepared using a 1D ligation sequencing kit (SQK-LSK-108) without fragmentation and sequenced using a MinION FLO-MIN-106 flow cell. Finally, ONT and Illumina short reads were combined to generate a hybrid assembly using Unicycler v0.4.4 (normal assembly mode, default parameters).

Phylogeny, recombination, and genomic sequence analyses

We built a core genome multiple sequence alignment (cg-MSA) from the assembled genome sequences. For this, the genome sequences were annotated using PROKKA v1.14.2 [52] with default parameters, resulting in GFF files. Roary v3.6 [53] was used to define protein-coding gene clusters, with a threshold set at 70% amino acid identity. Core genes were defined as being present in 95% of genomes and were concatenated into a cg-MSA by Roary. ClonalFrameML v1.11 [54] was used to build a phylogenetic tree based on the cg-MSA, which quantifies and accounts for the effects of recombination events. PhyML v20131022 [55] was used to build an initial tree. To evaluate bootstrap support, IQtree version 2 [56] with best-fit model GTR+F+R10 was used.

We used Kleborate v1.0.0-beta (https://github.com/katholt/Kleborate), with the --resistance option, to identify (identity > 80% and coverage > 90%) known resistance genes in *C. diphtheriae* genomic sequences, based on the August 1, 2019, update of the ARG-Annot database. We used BLASTN (identity > 80% and coverage > 95%) to search for the presence of the *tox* gene and of genes associated with biovar Gravis (DIP351, DIP354, and DIP357) and for nitrate utilization (*narKGHJI*).

MLST genotypes were defined using the international MLST scheme for *C. diphtheriae* and *C. ulcerans* [12].

Genome-wide association studies

The software treeWAS [57] was used to find genome-wide associations between either antimicrobial resistance phenotypes or biovar on the one hand and genetic variants (both core-genome SNPs and accessory genome gene presence/absence) on the other hand. Core-genome SNPs were derived either from a mapping approach (Samtools v1.9 and GATK v3.4-0), which comprises intergenic regions, or from the alignment of core coding sequences found using Roary. We ran treeWAS v1.1 with default parameters, using as input the previously computed ClonalFrameML phylogenetic tree and distribution of homoplasies, in order to account for both the population structure and effect of recombination. For this analysis, susceptibility phenotypes were classified into resistant or susceptible categories based on zone diameter phenotypes using the CA-SFM/EUCAST 2019 cutoffs (Fig. 3, Additional file 3: Table S3). The seven chromosomal PBP coding sequences (including gene with locus tag RS14485 in RefSeq NC_002935.2, or DIP0637 in the original GenBank file, renamed by us as PBP4b) were extracted from the genomic sequences and translated into amino acid (AA) sequences, which were also analyzed for association with penicillin resistance.

Cloning and transformation experiments

For ectopic expression in *C. glutamicum*, the *pbp2m* gene was amplified from *C. diphtheriae* strain FRC0402 and put under the control of the inducible P_{gntK} promoter on the shuttle vector pTGR5 [58] (Additional file 4: Fig. S1). *pbp2m* was assembled in this plasmid by Gibson assembly using the primers PBPdi_Fw (CAA AGA AAG GAT AAG ACC ATA TGA TGA CTA AGC ACA ATC GTT TCC GTC), PBPdi_Rv (TAC CTT AAG CGG CCG CTT TAT TGA ATT CCA GAG AAT TTC TGA ACA TCC G), pTGRdi_Fw (TAA AGC GGC CGC TTA AGG TAC C), and pTGRdi_Rv (ATG GTC TTA TCC TTT CTT TGG TGG CG).

Escherichia coli CopyCutter EPI400 (Lucigen) was used for the cloning of the *pbp2m* gene and was grown in Luria-Bertani (LB) broth or agar plates at 37 °C supplemented with 50 μg/mL kanamycin. The pTGR5_pbp2m plasmid was sequenced and electroporated into *Corynebacterium glutamicum* ATCC 13032. Positive colonies were grown in brain heart infusion (BHI) at 30 °C and 120 rpm supplemented with 25 μg/mL kanamycin and 1% (w/v) gluconate when required for ectopic expression of Pbp2m.

Mapping of SNPs on *C. diphtheriae* PBP sequences

Functional annotation of the different sequences was performed with InterPro [59]. Conserved transpeptidation motifs SxxK, SxN, and KTG were identified and mapped on the PBP sequences from *Corynebacteriales* based on the results of multiple sequence alignments performed with Clustal Omega [60]. When uncertainty between a transmembrane domain and a signal peptide existed, a decision was made based on the previous characterization of the homologous PBP in other *Corynebacteriales* in the literature.

Results

Provenance and microbiological characteristics of *C. diphtheriae* isolates

We studied 247 *C. diphtheriae* strains of diverse geographic and temporal origins (Fig. 1). This collection included 163 isolates prospectively collected between 2008 and 2017 from French mainland and overseas territories, 15 older (1981–2006) French clinical isolates, 65 ribotype reference strains [40], and 4 other reference strains. All isolates were confirmed as *C. diphtheriae* (excluding *C. belfantii* and *C. rouxii*) based on an average nucleotide identity (ANI) value higher than 96% with the *C. diphtheriae* type strain NCTC11397^{T}.

Approximately one third ($n = 78$, 32%) of isolates were *tox*-positive (as defined by the detection of the *tox* gene by PCR), whereas the remaining 169 isolates (68%) were *tox*-negative. The proportions of *tox*-positive isolates were 42%, 34%, and 2% among reference strains, 2008–2017 clinical isolates, and older clinical isolates, respectively (Additional file 4: Fig. S2). Of the 78 *tox*-positive isolates, 17 (21.8%) were negative for toxin production and thus correspond to non-toxigenic toxin gene-bearing (NTTB) isolates. Six of the NTTB isolates had a stop codon within the *tox* gene sequence (Additional file 1: Table S1, Additional file 5: Table S4). However, for the 11 remaining strains, we found no explanation for the observed lack of toxin production.

Upon biotyping, 154 (62.3%) isolates belonged to biovar Mitis, 87 to biovar Gravis (35.2%), and 6 (2.4%) to biovar Belfanti (Additional file 1: Table S1). Biovar proportions were similar among the three datasets. Mitis isolates were more frequently *tox*-positive than Gravis isolates (56/154 versus 18/87, chi-squared test, *p* value 0.01; Additional file 4: Fig. S2). Among *tox*-positive isolates, NTTB were more frequent among Mitis isolates (13/56, 23.2%) than among Gravis isolates (1/18, 5.6%) although this difference was not statistically significant (*p* value 0.09). Three out of four *tox*-positive Belfanti isolates were NTTB.

Phylogenetic structure of *C. diphtheriae* and distribution of the toxin gene

To infer a phylogenetic tree, we first aimed to detect and remove homologous recombination events among *C. diphtheriae* genomic sequences. ClonalFrameML inferred a relative rate of recombination to mutation (*R*/theta) of 0.86, with an average length of recombination segments (delta) of 287 bp. The mean genetic distance between donor and recipient of recombination (nu) was estimated by ClonalFrameML to be 0.02, resulting in a relative impact of recombination to mutation in genomic diversification (r/m = *R*/theta × delta × nu) of 5.01 [61].

The recombination-corrected phylogeny was similar to the uncorrected phylogeny but with shorter branches, as expected (Additional file 4: Fig. S3). The phylogeny, rooted with *C. belfantii* and *C. rouxii* (Additional file 4: Fig. S4), was star-like, with a multitude of sublineages branching off deeply. The deepest branching sublineages corresponded to two ribotype reference strains of biovar Mitis: CIP107521 (ribotype Dagestan) and CIP107534 (ribotype Kaliningrad). Remarkably, isolates of biovars Mitis and Gravis were mostly distributed in two distinct branches of the tree. We therefore named the two major branches, lineage Mitis (156 strains, of which 86% were of biovar Mitis) and lineage Gravis (91 strains, of which

77% were of biovar Gravis). The Gravis lineage branched off from within the Mitis lineage (Fig. 2; Additional file 4: Fig. S5). However, although most shallow branches were strongly supported, bootstrap support values of the deep nodes of the phylogeny were generally weak, implying that the position of the Gravis branch within the Mitis diversity was uncertain, as were the relative positions of the early-branching Gravis and Mitis sublineages (Additional file 4: Fig. S5, Fig. S6). Reference strains PW8 and NCTC11297^T belonged to the Mitis lineage, whereas NCTC13129 (from the ex-Soviet Union 1990's outbreak) and NCTC10648 belonged to the Gravis lineage. The Belfanti isolates were scattered in three distinct sublineages within the Mitis lineage and one within the Gravis lineage.

The isolates carrying the *tox* gene belonged mostly to the Mitis lineage (68 of 78, 87.2%), in which they were distributed in multiple sublineages. In the Mitis lineage, 69 (44.2%) were *tox*-positive. In contrast, within the Gravis lineage, only 10 (11%) isolates were *tox*-positive, and they corresponded to the earliest-branching Gravis sublineages with only one exception. Interestingly, this exception corresponded to the large ex-Soviet Union outbreak in the 1990s ((Fig. 2; Additional file 4: Fig. S5). This phylogenetic pattern is consistent with an evolutionary scenario where Mitis is the ancestral biovar of *C. diphtheriae* and where Gravis evolved from the Mitis lineage as an initially *tox*-positive sublineage, with subsequent loss of the toxin gene. In this scenario, the ex-Soviet Union outbreak sublineage would have reacquired the *tox* gene. All NTTB isolates belonged to the Mitis lineage except strain CIPA99 (ribotype Rhone, biovar Belfanti; Fig. 2), and they were distributed in multiple sublineages, showing convergent evolution towards the loss of toxin production.

Genetic events linked to biovar status

Biovar Mitis and Gravis are distinguished by the ability to utilize glycogen (positive in Gravis, negative in Mitis). The *spuA* gene, which codes for a putative alpha-1,6-glycosidase, was reported as being specific for biovar Gravis isolates [62]. Our genome-wide association study (GWAS) of accessory genes with the biovar phenotype revealed a strong association of a cluster of genes that includes *spuA* (DIP357; Additional file 4: Fig. S7) with biovar Gravis isolates, providing statistical support to the discovery of Santos et al. [62]. This association was stronger within the Gravis lineage; in contrast within the Mitis lineage, few (5 out of 17) of the biovar Gravis isolates possessed *spuA* (Fig. 2). GWAS analysis of core SNPs further demonstrated that a SNP (at position 324,487, Additional file 4: Fig. S7) downstream of the *spuA* cluster insertion point was also associated with biovar, suggesting homologous recombination among core genes as a mechanism for the *spuA* cluster insertion event.

The nitrate reductase activity differentiates Mitis and Gravis isolates, which are positive, from Belfanti isolates, which are nitrate-negative. We found that the nitrate reduction *narKGHJI* gene cluster [62] was disrupted in three of the six isolates assigned to the biovar Belfanti: strains FRC0480 and FRC0481 had a G to A mutation at position 675 of the *narG* gene, leading to a stop codon, whereas in strain CIPA99, approximately 100 nucleotides were inserted at position 446 in *narG*. No molecular explanation was found for the lack of nitrate reductase ability of the three other Belfanti strains when scrutinizing the *narKGHJI* gene cluster and adjacent molybdenum cofactor biosynthesis genes [63].

Antimicrobial susceptibility variation

Susceptibility to 19 antimicrobial agents was determined for the 247 clinical isolates and reference strains (Additional file 3: Table S3). For each agent, the distribution of zone diameter (ZD) values (Fig. 3) revealed a predominant mode located towards the right end of the distribution. This mode likely corresponds to the natural susceptibility distribution within the *C. diphtheriae* population and was used to define tentative epidemiological cutoffs (ECOFFs, also called ecological cutoff [64]). The proposed ECOFFs and their comparison with clinical breakpoints are presented in Additional file 2: Table S2. For each antimicrobial agent except cefotaxime, this approach led to the identification of strains that had ZD values smaller than those in the main mode, thus potentially corresponding to acquired resistance (Fig. 3).

Penicillin was exceptional in that the predominant mode (centered around 36 mm) was less neatly defined, due to the partial overlap with a second mode of smaller diameter values centered around 24 mm. This second mode corresponds mostly to the "intermediate" interpretative category ($18 \le ZD < 29$ mm) but also overlaps with the "resistant" category (< 18 mm). The distribution of ZD values for tetracycline also showed a clear second mode. For multiple other agents (amoxicillin, oxacillin, imipenem, kanamycin, rifampicin, ciprofloxacin, clindamycin and more evidently sulfonamide, trimethoprim, and the trimethoprim-sulfamethoxazole combination), some strains had the minimal diameter (6 mm, corresponding to growth at the disk contact). For trimethoprim, we observed both a second mode centered around 14 mm and a group of even more resistant strains with growth at disk contact.

Antimicrobial resistance levels were similarly distributed between *tox*-positive and *tox*-negative isolates (Additional file 4: Fig. S8) as well as between the two main phylogenetic lineages or biovars (Additional file 4: Fig. S9; Additional file 1: Table S1). In particular, the

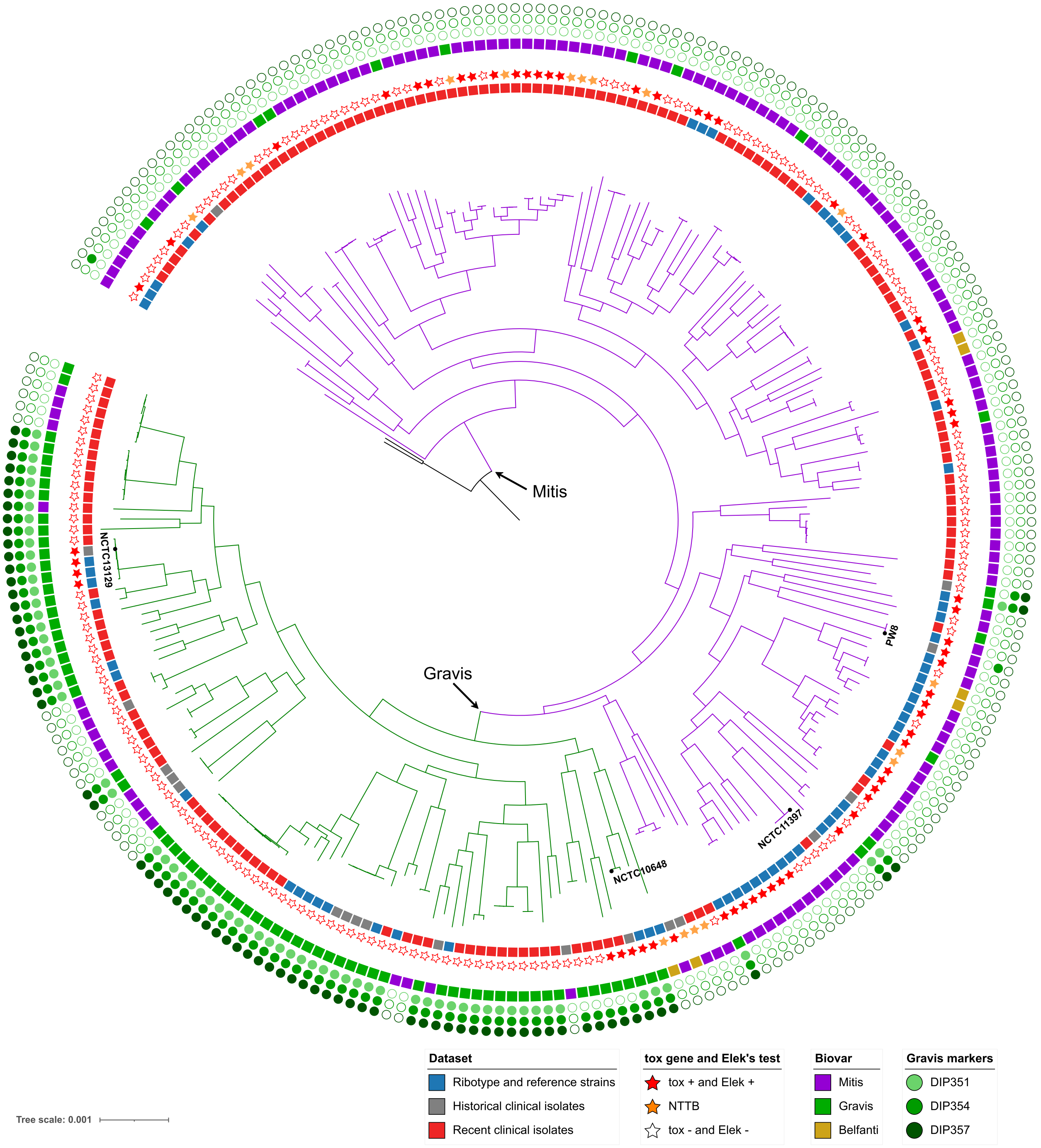

Mitis
Gravis
NCTC13129
PW8
NCTC11397
NCTC10648
Tree scale: 0.001
Dataset
Ribotype and reference strains
Historical clinical isolates
Recent clinical isolates
tox gene and Elek's test
tox + and Elek +
NTTB
tox - and Elek -
Biovar
Mitis
Gravis
Belfanti
Gravis markers
DIP351
DIP354
DIP357

Fig. 2 Phylogenetic tree of *C. diphtheriae*. The tree was obtained using ClonalFrameML and was rooted using *C. rouxii* and *C. belfantii* isolates (see supplementary material). The main lineages Mitis and Gravis are labeled, and their branches are drawn using purple and green, respectively. The first (internal) circle around the tree corresponds to the three strain subsets (red, recent clinical isolates; blue, reference strains; gray, older clinical isolates). The second circle (stars) gives the toxigenic status. The third circle corresponds to biovars Mitis (purple), Gravis (green), and Belfanti (yellow). The next three circles indicate the presence of the spuA-associated gene cluster; DIP357 = *spuA* gene. The positions of reference strains PW8, NCTC13129, and NCTC10648 are indicated. The scale bar gives the number of nucleotide substitutions per site

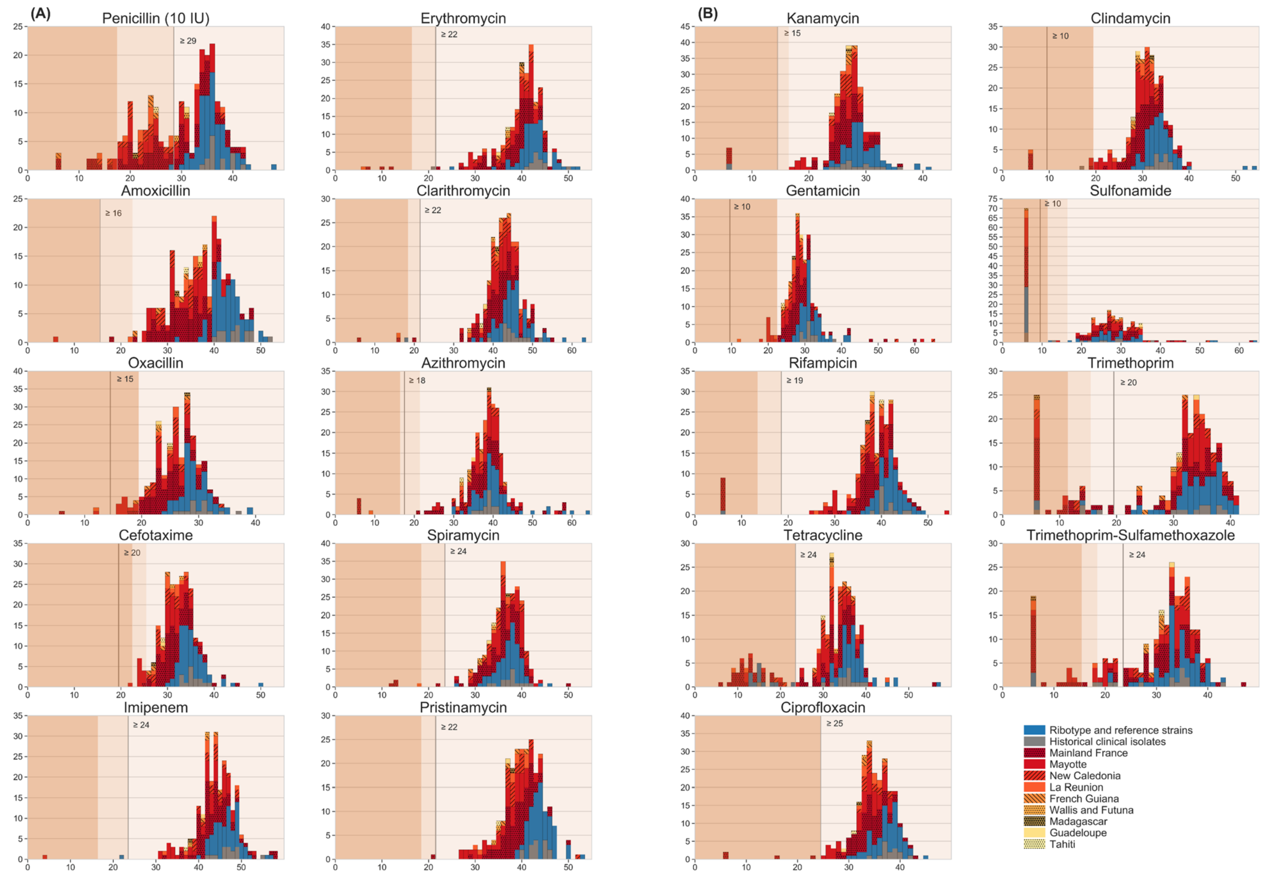


Fig. 3 The distributions of zone diameter values for 19 antimicrobial agents. **a** Beta-lactams, macrolides, and pristinamycin. **b** Other agents. *X*-axis: diameter in millimeters; *Y*-axis: number of strains. The colors inside the bars represent the subset and geographic origins as in Fig. 1 (see key on **b**). The three background colors represent the categorical interpretations according to EUCAST: resistant (salmon, left), intermediate (lighter salmon, middle), and susceptible (pale beige, right). The gray vertical bar corresponds to the proposed tentative ecological cutoff

proportion of multidrug-resistant strains was not statistically different between biovars Mitis ($n = 14$) and Gravis ($n = 4$; chi-square test p value 0.2) or between *tox*-positive and tox-negative isolates (6 versus 12, p value 0.87).

Resistance rates were 17.2%, 2.5%, and 2.5% for penicillin, amoxicillin, and erythromycin, respectively, among the prospectively collected 2008–2017 clinical isolates (Fig. 4). Reference strains were generally susceptible to most agents, including penicillin, but were partially resistant to tetracycline (18%) and sulfonamide (35%). The resistance profile distribution showed that approximately half (121/247) of the strains had a fully susceptible phenotype, whereas 18 isolates (7.3%) were multidrug-resistant (Table S1). Notably, four isolates were resistant at the same time to penicillin and macrolides (Fig. 4 inset), and two of them (FRC0402 and FRC0466) additionally had a reduced susceptibility to amoxicillin. Two of these isolates (FRC0475 and FRC0478) were collected from a foot arch wound and in respiratory carriage in the same patient (French mainland, with recent travel from New Caledonia). The two others came from a patient living in La Réunion Island (FRC0402) and from a patient living in Paris, who had recently traveled to Tunisia (FRC0466). Of the 18 multidrug-resistant isolates, 1 was a 1994 isolate from India and 17 belonged to the 2008–1017 prospective subset of clinical isolates (representing 10.4% of these). Isolates from La Réunion Island and mainland France showed resistance to multiple antimicrobial agents more often than isolates from other geographic origins (Fig. 1a, inset).

Genomic associations with antimicrobial resistance phenotypes

We first searched for the presence in the genomic sequences, of previously described antimicrobial resistance genes (ARGs). This approach led to the detection of 12 ARGs (Additional file 6: Table S5). We identified three tetracycline resistance genes (*tetW*, *tet33*, and *tetO*), four aminoglycoside resistance genes [aph (3′)-Ia, aph (3″)-Ib, aph (6)-Id, and aadA1 = ant (3′)-Ia], and also *ermX*, *dfrA16*, *dfrA15b*, *dfrA1*, and *sul1* genes. We observed a strong correlation between the presence of ARGs and the

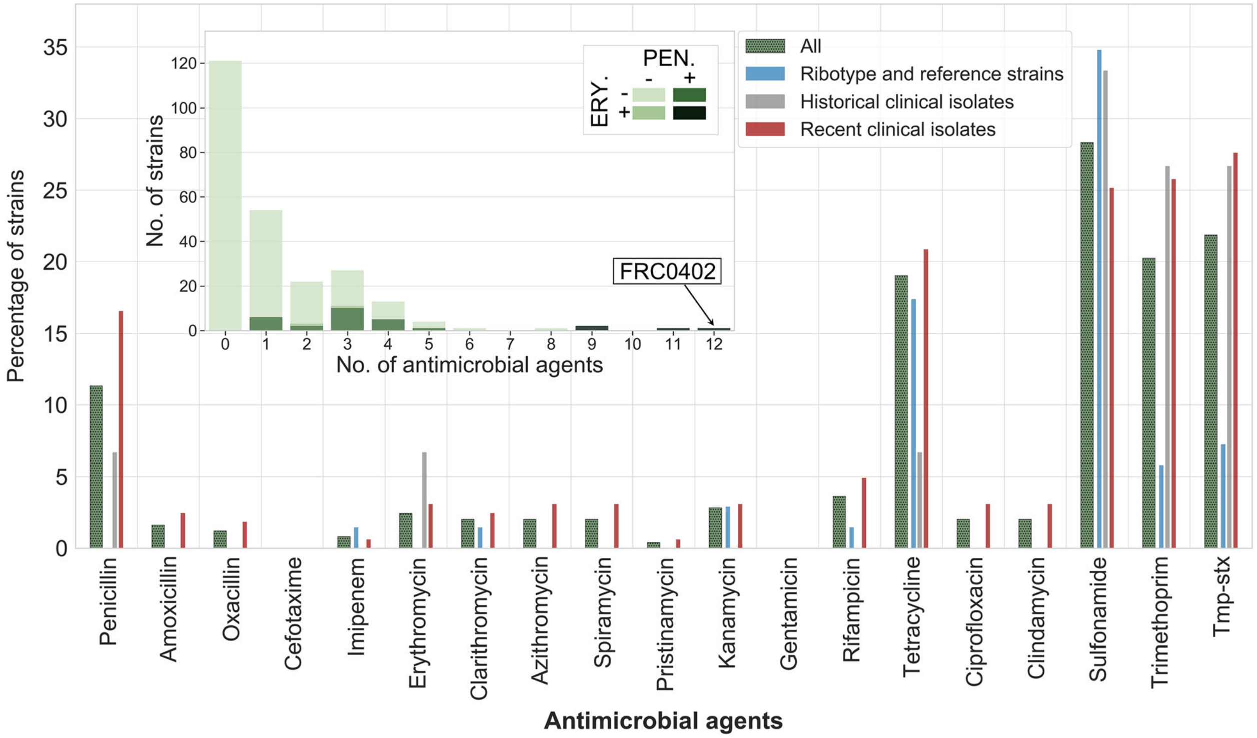


Fig. 4 Proportions of resistant strains by antimicrobial agent and multidrug resistance distribution. Interpretation of the zone diameter values was performed according to the proposed ecological cutoffs. The main panel shows the percentage of strains resistant to each agent. Tmp-stx, trimethoprim-sulfamethoxazole. The four bars for each agent correspond to the entire dataset (all, shaded green) or the three subsets (see key). The inset shows the number of strains resistant to a given number of antimicrobials. Penicillin- and/or erythromycin-resistant strains are colored with darker green (see key). FRC0402, the most multidrug-resistant isolate with resistance to 12 agents, is highlighted

expected resistance phenotypes (Additional file 4: Fig. S10), particularly for *ermX* (macrolide resistance; phi coefficient = 0.66 with erythromycin resistance), *sul1* (sulfonamide resistance; phi = 0.80), and *aph (3′)-Ia* (kanamycin resistance; phi = 0.84). In strains FRC0137 and FRC0375, this latter gene was linked to *aph (3′)-Ib* (strA), *aph (6)-Id* (strB), and *ermX* on a Tn5432-like genomic region with an IS1249 insertion sequence [65]. The phylogenetic distribution of ARGs (Fig. 5; Additional file 4: Fig. S6) revealed their presence in multiple unrelated sublineages, consistent with independent acquisitions by horizontal gene transfer. Gene *ermX* was present either in proximity to gene *pbp2m* (see below) or in a fragmented insertion-sequence rich accessory region. Gene *dfrA16* was associated with *sul1* on a reported [37] class 1 integron (see the "Discovery of a multidrug-resistant conjugative plasmid carrying the gene *pbp2m*" section). Tetracycline resistance was associated either with the ribosomal protection protein genes tet(O) or tet(W) or with the efflux pump gene tet33. These three genes were present in distinct strain subsets and appear to contribute independently to tetracycline resistance in *C. diphtheriae* (Fig. 5); they were mostly associated with insertion sequences but not with other ARGs.

Next, in order to identify novel genetic determinants potentially associated with antimicrobial resistance in *C. diphtheriae*, a GWAS approach was followed, based on either core genome SNPs or accessory gene presence/absence. SNPs that were strongly associated with ciprofloxacin, trimethoprim, and rifampicin resistance were identified within genes for gyrase subunit A, dihydrofolate reductase, and RNA polymerase subunit B, respectively (Additional file 6: Table S5; Additional file 4: Fig. S11), consistent with known mechanisms and validating our approach. SNPs were also found to be associated with penicillin, kanamycin, and tetracycline resistance (Additional file 6: Table S5), but based on the annotations of the genes where they are located, the functional implications of these SNPs are unclear in these cases. No association was found for penicillin resistance within the core PBP genes using the genome-wide approach. However, using a concatenation of the amino acid sequences of the seven identified putative PBPs of *C. diphtheriae*, we identified amino acid positions that were statistically associated with penicillin resistance (Additional file 4: Fig. S12). The identified positions were mapped onto the predicted functional domains of the different PBPs (Additional file 4: Fig. S13) revealing several mutation hotspots. A number of significant SNPs were

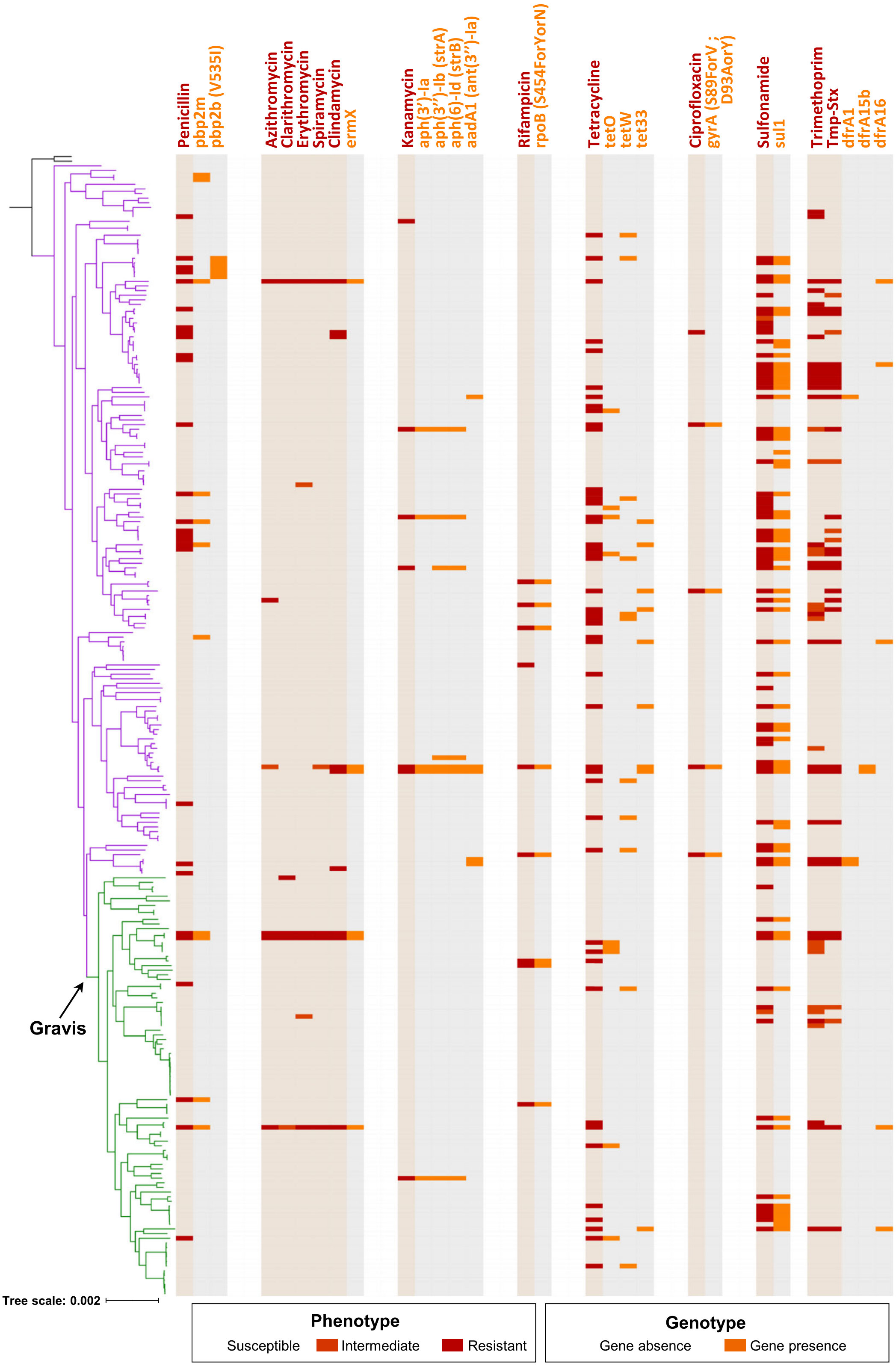


Fig. 5 (See legend on next page.)

(See figure on previous page.)
Fig. 5 Phylogenetic distribution of antimicrobial resistance phenotypes and genes. The phylogenetic tree is the same as in Fig. 1; the Mitis branch is in purple, the Gravis branch in green. To the right of the tree, each bloc indicates first the phenotype (red, resistant; see key) and relevant corresponding genotypes (orange, gene or mutation presence)

found within the transpeptidase (TP) domains of the different PBPs, but none of the mutations mapped to the conserved transpeptidation motif. Other mutations were found outside the TP domains, for instance, in the transglycosylase and PASTA domains of PBP1b or in the dimerization domain of PBP2b.

GWAS analysis of accessory genes demonstrated significant associations with phenotypic resistance. Associated genes included those mentioned above for erythromycin, tetracycline, kanamycin, sulfonamide, and trimethoprim (Additional file 6: Table S5), showing that these genes are the main mechanisms of resistance to these antimicrobial agents in *C. diphtheriae* populations. In addition, an accessory penicillin-binding protein gene (which we name *pbp2m*) was strongly associated with penicillin resistance. This gene is described in more detail below.

Discovery of a penicillin-binding protein (PBP2m) associated with penicillin resistance

The PBP gene *pbp2m* was observed in 11 isolates, 8 of which were penicillin resistant with minimum inhibitory concentrations (MIC) ranging from 0.19 to 1.5 mg/L. Two of these isolates were also resistant to amoxicillin and one was in addition resistant to oxacillin (Additional file 2: Table S2). The phylogenetic distribution of *pbp2m*-positive strains was compatible with multiple independent acquisitions of the gene through horizontal gene transfer (Fig. 5). Sequence analysis showed that the newly identified PBP2m is almost identical (3 differences out of 593 amino acids, 99.5%) to PBP2c from *C. jeikeium*, a class B PBP with an N-terminal signal peptide followed by a lipobox domain and the C-terminal transpeptidase domain (Additional file 4: Fig. S13). The *C. jeikeium* PBP2c is a low-affinity PBP and was associated with beta-lactam resistance in *C. jeikeium* [66].

To demonstrate the role of PBP2m in penicillin resistance, its gene was PCR amplified from FRC0402 and cloned into the pTGR5 plasmid (Additional file 4: Fig. S1). Transformation of the plasmid into *C. glutamicum* strain ATCC 13032 raised the MIC for penicillin from 0.125 to 1.5 mg/L, and the MICs of the other beta-lactams amoxicillin, oxacillin, and cefotaxime also increased importantly (Fig. 6; Additional file 7: Table S6). In contrast, MICs of non-beta-lactam agents were not changed. Imipenem was less effective against the transformant based on disk diffusion but not based on E-test. Transformation with the empty plasmid used as control

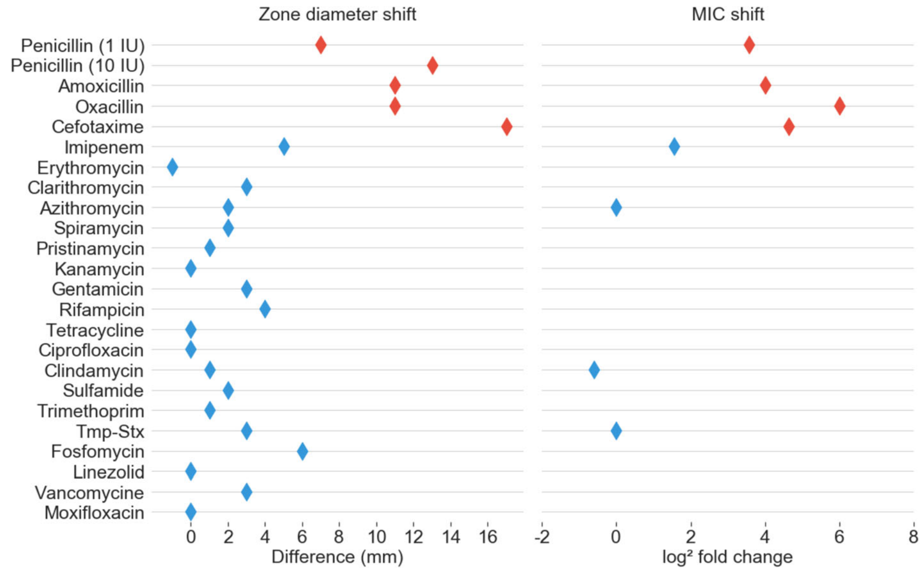


Fig. 6 Phenotypic effect of *pbp2m* expression. Compared susceptibility phenotypes for *C. glutamicum* transformants with plasmid pTGR5 containing, or not, the *pbp2m* gene. Left, the shift in zone diameter size; right, the shift in the minimum inhibitory concentration (MIC). Diamonds are positioned on the scales, at positions corresponding to the difference of zone diameters (without *pbp2m* – with *pbp2m*) or the log2 of MIC ratios (with *pbp2m*/without *pbp2m*). Red, penicillins or cephalosporins; blue, other agents. Tmp-Stx, trimethoprim-sulfamethoxazole

did not affect the MIC of any agent. These results show that PBP2m confers resistance to a broad range of beta-lactams.

Discovery of a multidrug-resistant conjugative plasmid carrying the gene *pbp2m*

Strain FRC0402, a *tox*-negative isolate from La Réunion Island, stood out as being resistant to 12 agents (Fig. 4 inset). In addition to *pbp2m*, this isolate carried genes *sul1*, *ermX*, and *dfrA16* and a *tetA* family *tet(Z)*-like (71%) tetracycline efflux gene. To define the genomic context of resistance genes, a complete genome sequence was obtained. The assembly revealed a chromosome of 2,397,465 bp and a circular plasmid of 73,763 bp (Fig. 7), which we propose to name pLRPD (for large resistance plasmid of *C. diphtheriae*).

The *pbp2m* gene was located on the large plasmid in a region comprising three other genes, a *blaB* beta-lactamase family gene, a LysR family regulator gene, and the *ermX* gene, flanked by two insertion sequences (IS*1628*) of the IS6 family (Fig. 7). With disparate direct repeat sequences, it remains unclear if this region

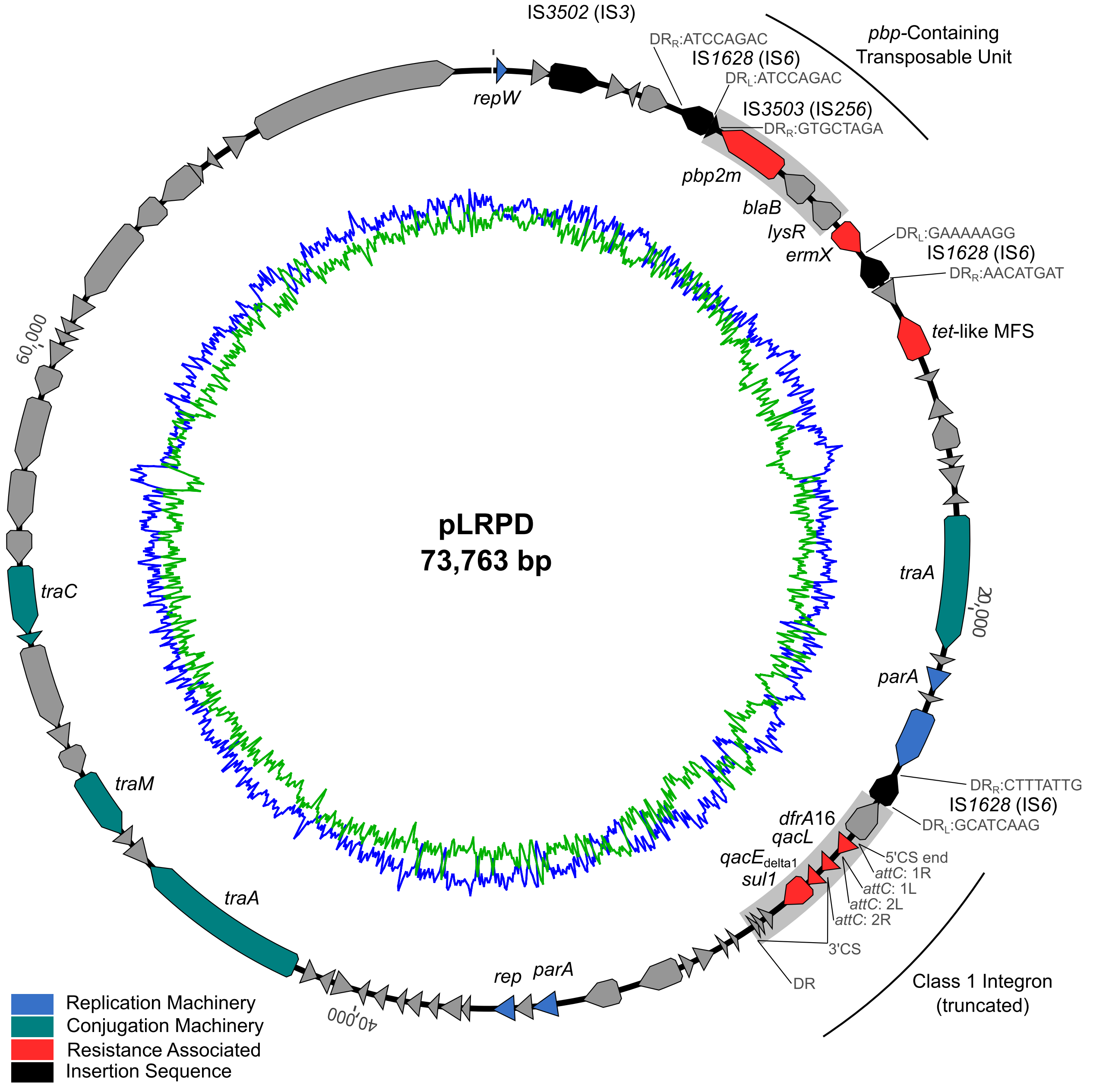

IS3502 (IS3)
DR$_R$:ATCCAGAC
IS1628 (IS6)
DR$_L$:ATCCAGAC
pbp-Containing Transposable Unit
IS3503 (IS256)
DR$_R$:GTGCTAGA
repW
pbp2m
blaB
lysR
ermX
DR$_L$:GAAAAAGG
IS1628 (IS6)
DR$_R$:AACATGAT
tet-like MFS
60,000
pLRPD
73,763 bp
traC
traA
20,000
parA
DR$_R$:CTTTATTG
IS1628 (IS6)
DR$_L$:GCATCAAG
dfrA16
qacL
qacE$_{delta1}$
sul1
5'CS end
attC: 1R
attC: 1L
attC: 2L
attC: 2R
3'CS
DR
Class 1 Integron (truncated)
traM
traA
rep
parA
40,000
Replication Machinery
Conjugation Machinery
Resistance Associated
Insertion Sequence

Fig. 7 Map of plasmid pLRPD from isolate FRC0402. Predicted coding sequences are portrayed by arrows and colored based on the predicted gene function (refer to key). Inner blue circle, G+C% content; inner green circle, A+T% content. IS3502 annotation is putative; IS3503 is truncated

represents a single transposable unit or a mosaic of gene acquisition events in pLRPD. A nearly identical PBP was observed in 78% of publicly available *C. jeikeium* genomes, in 57% of *C. striatum* genomes and in multiple other *Corynebacterium* genomes (Additional file 8: Table S7). However, the genetic context of PBP2m was highly variable in *C. diphtheriae* and among other *Corynebacterium* species (Additional file 4: Fig. S14). A putative transposable PBP-containing unit (PCU) comprising genes *pbp2m*, *blaB*, and *lysR*, commonly flanked by IS*3503* (IS*256* family) with a fragment identified in pLRPD (Fig. 7), appeared to be highly conserved and was associated variably with *ermX* in *C. diphtheriae* and with a helicase in *C. diphtheriae* and other *Corynebacterium* species. The PCU was sometimes found in 2 or 3 tandem copies and was chromosomally located in most genomes.

Further elements carried by pLRPD included an integron carrying genes *dfrA16*, *qacL*, and *sul1*, as well as elements of a putative conjugation apparatus gene cluster (Fig. 7). Our conjugation experiments aiming to demonstrate the transfer of pLRPD into recipient *C. diphtheriae* isolates failed. This plasmid was not found in other *C. diphtheriae* strains.

Discussion

Strains of *C. diphtheriae* that are resistant to antimicrobial therapy may compromise the management of diphtheria cases and the control of pathogen transmission. Here, we aimed to define the genomic determinants of resistance to penicillin and other antimicrobial agents in *C. diphtheriae* and to analyze the relationships of resistance with diphtheria toxin production, biochemical variants, and phylogenetic sublineages. To this aim, we characterized phenotypically and genotypically a large sample of *C. diphtheriae* isolates from diverse geographic and temporal origins. We confirmed that the species is made of multiple phylogenetic sublineages [9, 13, 67] and showed that homologous recombination contributes five times more to their diversification than mutation, consistent with previous evidence of recombination in *C. diphtheriae* populations [12, 13].

Historically, *C. diphtheriae* isolates have been classified into three main biovars, but the links between biovars and phylogenetic structure have remained obscure. Whereas previous work concluded on the absence of an association [11], our phylogenetic analyses reveal that Gravis and Mitis, the two main biovars of *C. diphtheriae*, are associated strongly with two phylogenetic lineages. Lineage Gravis appears to have acquired ancestrally a gene cluster comprising the extracellular glycogen debranching enzyme gene *spuA* [62]. Although the Gravis phenotype was associated with *spuA* within lineage Gravis, other genomic determinants of glycogen utilization remain to be discovered within the Mitis lineage. Whereas most of our biovar Belfanti isolates were excluded from this work because they belonged to *C. belfantii* or *C. rouxii*, a few Belfanti isolates did belong to *C. diphtheriae.* For three of them, mutations leading to a non-functional nitrate reductase were uncovered, explaining the phenotypic switch to biovar Belfanti. Our results show that biotyping is subject to parallel evolution, perhaps due to a yet-unknown selective pressure, and has limited species-level identification (as Belfanti is found in the three species *C. diphtheriae*, *C. belfantii*, and *C. rouxii*) or epidemiological typing value.

The most important factor of *C. diphtheriae* pathogenicity is the diphtheria toxin. Despite early realization that it is encoded on a prophage [5], few studies have investigated the phylogenetic distribution of the *tox* gene in *C. diphtheriae* [8, 9]. Here, we show that *tox*-positive strains mainly belong to the Mitis lineage and to early-diverging branches of the Gravis lineage. The distribution of *tox*-positive isolates into multiple Mitis sublineages is strongly indicative of independent acquisitions of the toxin gene. Alternately, this pattern might result from initial acquisition of the *tox* gene, followed by secondary loss in multiple sublineages. The phylogenetic pattern is also consistent with an ancestral presence of the *tox*-bearing phage in the Gravis lineage, with subsequent loss of the *tox* gene in the branch leading to the ancestor of most Gravis isolates. Future work should investigate the dynamics of the lysogenic corynephages and molecular determinants of their sublineage distribution. One important open question is the likelihood of *tox*-negative strains acquiring the *tox* gene during colonization, infection, or short-term epidemiological timeframes [7].

Remarkably, except for early-diverging sublineages, only one Gravis sublineage was found to carry the *tox* gene. This sublineage happens to correspond to the largest outbreak in recent times, which occurred in Newly Independent States of the ex-Soviet Union in the 1990s [13, 14, 38]. This sublineage, which comprises the ST8 reference strain NCTC13129, is genetically distant from other *tox*-positive lineages, which belong to the Mitis lineage. Hence, its antigenic structures or other pathogenicity properties may have diverged from those of more common *tox*-positive isolates, which might have contributed to its exceptional transmission in the 1990s, in addition to the decline in vaccine coverage [14]. Of note, biovar Gravis was named to reflect a perceived higher severity of infection compared to diphtheria cases caused by biovar Mitis isolates [68, 69]. Recently, it was shown that most diphtheria vaccines contain, besides the anatoxin, multiple other *C. diphtheriae* immunogens [22]. The impact of vaccination on the evolution of *C.*

diphtheriae populations, and possible variations of cross-protection as a function of strain diversity, is currently undefined. This work provides a framework onto which future studies can build to address this important question.

Although antimicrobial-resistant *C. diphtheriae* strains have been reported on numerous occasions [23, 32], knowledge on antimicrobial resistance in *C. diphtheriae* is largely fragmented and suffers from the lack of harmonization. Breakpoints used to define resistance vary according to world region and have changed over time within single countries [70]. The lack of consensus on the definition of resistance restricts our ability to define the magnitude of the problem and its global significance.

We aimed to define biologically meaningful cutoffs [64] based on susceptibility phenotype distributions, taking advantage of our large and diverse sample. Our data allowed us to propose tentative ecological cutoffs for *C. diphtheriae*. This approach should in the future be extended to MIC values [71] and should use larger and more diverse strain collections. Nevertheless, our analyses suggest that non-susceptibility to at least one antimicrobial agent was acquired by half of *C. diphtheriae* strains, regardless of lineage, biovar, or toxigenic status. This study further suggests that acquired resistance to penicillin, the first-line therapy against diphtheria, is far from being rare, affecting > 15% of *C. diphtheriae* isolates collected in the last decade in France and its overseas territories. Besides, 10.4% of these isolates were multidrug-resistant. The high prevalence of resistance to penicillin, tetracycline, and trimethoprim/sulfamethoxazole found here is consistent with susceptibility surveys of recent *C. diphtheriae* isolates in Algeria [29], Indonesia [31, 72], and India [73]. Many high-income countries such as France have chosen to use amoxicillin as the first choice for antibiotic therapy [74], as this molecule remains highly active. Still, widespread penicillin resistance is concerning, since diphtheria mainly occurs in resource-poor settings where penicillin G is largely used. In contrast, resistance to erythromycin and other macrolides remains rare. Our results call for concerted research into the magnitude of the antimicrobial resistance threat in *C. diphtheriae*.

Knowledge of the genetic mechanisms of antimicrobial resistance is critical for defining appropriate treatments, refining diagnostics, and conducting epidemiological studies of antimicrobial resistance. Resistance genes to several antimicrobial classes have been described in *C. diphtheriae* [36, 37, 75], while additional genes described in other *Corynebacterium* species [76] might also be present in *C. diphtheriae*. Here, we defined the prevalence and phylogenetic distribution of previously reported and newly identified resistance determinants in *C. diphtheriae*. We demonstrate the co-occurrence of resistance phenotypes and genes, suggesting a causative link in multiple instances. We further show that resistance genes have been acquired independently in multiple sublineages, demonstrating a dynamic resistome in *C. diphtheriae*. In addition, we demonstrate an association between alterations in chromosomally encoded targets and phenotypic resistance for quinolone, trimethoprim, and rifampicin. Fluoroquinolone resistance was previously linked to mutations in the *gyrA* gene in *C. amycolatum* [77], *C. striatum* [78], and *C. belfantii* [79] but seemingly never for *C. diphtheriae*. Finally, we demonstrate the co-occurrence within some strains of multiple resistance determinants and uncover a previously undescribed large resistance plasmid in *C. diphtheriae*. The mechanism of genetic transfer of this plasmid remains to be investigated. This work provides the first overview of the *C. diphtheriae* resistome and will facilitate further studies into the evolutionary emergence of multiresistant *C. diphtheriae* strains.

Here, we discovered an accessory PBP (PBP2m) and experimentally showed it to confer resistance to penicillin and other beta-lactam antimicrobial agents. Its distribution in multiple sublineages, and its presence in other *Corynebacterium* species, clearly demonstrates its horizontal transfer, and we revealed a multiplicity of genomic contexts in which it is found within *Corynebacterium*. PBP2m is a putative low-affinity PBP, which would explain why it is less affected by beta-lactam antibiotics. Very recently, Forde et al. [80] have reported high-level penicillin resistance (MIC > 256 g/L) in *C. diphtheriae* strain BQ11 and have attributed this phenotype to a penicillin-binding protein located on transposon Tn3503. We found that the *pbp2m*-containing unit is 99.92% similar to Tn3503, which additionally harbors a helicase (Additional file 4: Fig. S14). Besides, PBP2m is 100% identical to the PBP from BQ11 and therefore also possesses the V361A alteration (using *C. jeikeium* as reference) that was hypothesized to confer high-level penicillin resistance [80]. Further studies on the expression of *pbp2m* in its variable genetic backgrounds are required to understand the observed difference in penicillin resistance levels. The location of *pbp2m* on a mobile element and its association with beta-lactam and carbapenem resistance represents a public health threat.

The seven chromosomal PBPs of *C. diphtheriae* (including the newly annotated PBP4b) were investigated to identify amino acid sequence polymorphisms associated with penicillin resistance. Although several alterations were significantly associated, none was directly implicated with the catalytic residues of the transpeptidase or transglycosylase domains. The association with resistance may not be directly linked to these catalytic residues but could be due to secondary sites that are thought to interfere with beta-lactam ligand binding. While biochemical and structural studies are necessary to understand how these mutations affect penicillin susceptibility, we postulate that some changes in these domains could lead to allosteric effects ultimately resulting in beta-lactam resistance, as described for *Staphylococcus aureus* PBP2a [81, 82] or *Streptococcus pneumoniae* PBP2x [83]. Other SNPs might simply have been hitchhiking due to their physical linkage with functionally important SNPs [84].

Conclusions

Our phylogenetic analyses showed that *C. diphtheriae* is highly diverse and comprises multiple sublineages that are grouped into two major lineages, characterized by distinctive associations with biovars Mitis and Gravis and with the diphtheria toxin gene. Antimicrobial resistance remains rare in France and its overseas territories, but multidrug-resistant strains were found and have a wide phylogenetic distribution. Analyses of the genetic bases of antimicrobial resistance defined the contribution of previously described mechanisms of resistance and uncovered a penicillin-binding protein, PBP2m, associated at the population level with penicillin resistance and shown experimentally to confer resistance to several beta-lactam antimicrobial agents. PBP2m is encoded in diverse genetic contexts, including on a large putatively conjugative resistance plasmid. This study reveals how biovars, diphtheria toxin gene, and antimicrobial resistance are associated with the phylogenetic diversity of *C. diphtheriae* strains, and provides novel data into the evolutionary dynamics of these medically important features. This framework will facilitate future studies investigating the emergence of sublineages of *C. diphtheriae* involved in diphtheria outbreaks, antimicrobial resistant infections, and the patterns and drivers of global spread of *C. diphtheriae* sublineages.

Supplementary Information

Supplementary information accompanies this paper at https://doi.org/10.1186/s13073-020-00805-7.

Additional file 1: Table S1. Strains used in this study and their characteristics.

Additional file 2: Table S2. Antimicrobial agents and zone diameters interpretative breakpoints used.

Additional file 3: Table S3. Susceptibility data and interpretative categories based on current EUCAST recommendations (SIR) and proposed ecological cutoffs (ECOFF).

Additional file 4. Contains Supplementary Figures S1 to S14.

Additional file 5: Table S4. Molecular explanations for non-toxigenic *tox* gene-bearing (NTTB) strains.

Additional file 6: Table S5. Associations with resistance phenotypes based on treeWAS results for accessory genes, SNPs in core genes and whole-genome SNPs; or based on Kleborate or direct analysis of potential target sequences.

Additional file 7: Table S6. Susceptibility phenotypes of transformed *C. glutamicum* ATCC 13032 strains.

Additional file 8: Table S7. Detection of the pbp2m gene in other *Corynebacterium* genomes.

Acknowledgements

We thank Vincent Enouf and the P2M core facility of Institut Pasteur for genomic sequencing. We are indebted to Collection de l'Institut Pasteur for providing reference strains of ribotypes, which were deposited following their described by Grimont and colleagues in 2004 [40]. We thank Valerie Bouchez for help with the Oxford Nanopore Technologies sequencing.

Authors' contributions

S.B. conceived, designed, and coordinated the study. A.C-L., E.B., M.B., and M.D. performed the microbiological cultures of the isolates and their biochemical and molecular characterizations. M.H., L.G.P., C.R., S.L.B., M.B.-P., M.D., and S.B. analyzed the genomic and phenotypic data. J.T. reviewed the clinical source data of the isolates. Q.G. and A.-M.W. performed the pbp2m cloning experiments and analyzed the PBP sequences. X.D. provided help with the phylogenetic, recombination, and GWAS analyses. S.B. wrote the initial version of the manuscript. All authors provided input to the manuscript. All authors read and approved the final manuscript.

Funding

MH was supported financially by a PhD grant from the European Joint Programme One Health, which has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 773830. LGP was supported financially by a grant from the Institut Français de Bioinformatique, the national infrastructure for services in bioinformatics created in the frame of the French Government's Investissement d'Avenir program. QG was funded by MTCI PhD school (ED 563). CR was supported by a Pasteur-Roux fellowship from Institut Pasteur. SLB received support from a Victorian Fellowship, provided by VESKI and funded by the State Government of Victoria, Australia.
The National Reference Center for Corynebacteria of the *Diphtheriae* complex receives support from Institut Pasteur and Public Health France (Santé Publique France, Saint Maurice, France). This work was supported financially by the French government's Investissement d'Avenir program Laboratoire d'Excellence "Integrative Biology of Emerging Infectious Diseases" (ANR-10-LABX-62-IBEID). AW and QG receive support from Institut Pasteur and the CNRS (France).

Availability of data and materials

The genomic sequencing data generated in this study were deposited in the European Nucleotide Archive (ENA) database and are accessible through the BioProject PRJEB22103 (https://www.ebi.ac.uk/ena/browser/view/PRJEB22103) [85].

Ethics approval and consent to participate

To conduct the research, we used bacterial strains, which are not considered human samples. Accordingly, this research was not considered human research and is out of the scope of the decree n° 2016-1537 of November 16, 2016, implementing law n° 2012-300 of March 5, 2012, on research involving human subjects. Therefore, no ethics approval was needed and no informed consent was required.
All French bacteriological samples were collected in the frame of the French national diphtheria surveillance and are collected, coded, shipped, managed, and analyzed according to the French National Reference Center protocols that received approval from the French competent body CNIL (Commission Nationale Informatique et Libertés; approval number: 1474671). Other strains were obtained from culture collections.
The research conformed to the principles of the Helsinki Declaration.

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests.

Author details
[1]Institut Pasteur, Biodiversity and Epidemiology of Bacterial Pathogens, Paris, France. [2]Collège doctoral, Sorbonne Université, F-75005 Paris, France. [3]Institut Français de Bioinformatique, CNRS UMS 3601, Evry, France. [4]Unité de Microbiologie Structurale, Institut Pasteur, CNRS UMR 3528, Université de Paris, F-75015 Paris, France. [5]Doherty Applied Microbial Genomics, Department of Microbiology & Immunology, The University of Melbourne at The Peter Doherty Institute for Infection & Immunity, Melbourne, Victoria, Australia. [6]Institut Pasteur, National Reference Center for Corynebacteria of the Diphtheriae Complex, Paris, France. [7]School of Life Sciences and Department of Statistics, University of Warwick, Coventry, UK. [8]Department of General Pediatrics and Pediatric Infectious Diseases, Hôpital Necker-Enfants Malades, APHP, Université de Paris, Paris, France.

Received: 29 May 2020 Accepted: 11 November 2020
Published online: 27 November 2020

References

1. Patey O, et al. Clinical and molecular study of Corynebacterium diphtheriae systemic infections in France. Coryne Study Group. J Clin Microbiol. 1997;35:441–5.
2. Hadfield TL, McEvoy P, Polotsky Y, Tzinserling VA, Yakovlev AA. The pathology of diphtheria. J Infect Dis. 2000;181(Suppl 1):S116–20.
3. Burkovski, A. *Corynebacterium diphtheriae* and related toxigenic species: genomics, pathogenicity and applications. (Springer, 2014).
4. Sharma NC, et al. Diphtheria. Nat Rev Dis Primer. 2019;5:81.
5. Freeman VJ. Studies on the virulence of bacteriophage-infected strains of Corynebacterium diphtheriae. J Bacteriol. 1951;61:675–88.
6. Parveen S, Bishai WR, Murphy JR. Corynebacterium diphtheriae: diphtheria toxin, the tox Operon, and its regulation by Fe2+ activation of apo-DtxR. Microbiol Spectr. 2019;7(4):GPP3-0063-2019.
7. Pappenheimer AM, Murphy JR. Studies on the molecular epidemiology of diphtheria. Lancet Lond. Engl. 1983;2:923–6.
8. Trost E, et al. Pangenomic study of Corynebacterium diphtheriae that provides insights into the genomic diversity of pathogenic isolates from cases of classical diphtheria, endocarditis, and pneumonia. J Bacteriol. 2012;194:3199–215.
9. Sangal V, Hoskisson PA. Evolution, epidemiology and diversity of Corynebacterium diphtheriae: new perspectives on an old foe. Infect. Genet. Evol. J. Mol. Epidemiol. Evol. Genet. Infect. Dis. 2016;43:364–70.
10. Timms VJ, Nguyen T, Crighton T, Yuen M, Sintchenko V. Genome-wide comparison of Corynebacterium diphtheriae isolates from Australia identifies differences in the Pan-genomes between respiratory and cutaneous strains. BMC Genomics. 2018;19:869.
11. Sangal V, et al. A lack of genetic basis for biovar differentiation in clinically important Corynebacterium diphtheriae from whole genome sequencing. Infect Genet Evol J Mol Epidemiol Evol Genet Infect Dis. 2014;21:54–7.
12. Bolt F, et al. Multilocus sequence typing identifies evidence for recombination and two distinct lineages of Corynebacterium diphtheriae. J Clin Microbiol. 2010;48:4177–85.
13. Grosse-Kock S, et al. Genomic analysis of endemic clones of toxigenic and non-toxigenic Corynebacterium diphtheriae in Belarus during and after the major epidemic in 1990s. BMC Genomics. 2017;18:873.
14. Dittmann S, et al. Successful control of epidemic diphtheria in the states of the former Union of Soviet Socialist Republics: lessons learned. J Infect Dis. 2000;181(Suppl 1):S10–22.
15. WHO. Surveillance and burden of diphtheria. http://www.who.int/immunization/monitoring_surveillance/burden/diphtheria/en/. (2018).
16. Galazka A. The changing epidemiology of diphtheria in the vaccine era. J Infect Dis. 2000;181(Suppl 1):S2–9.
17. Rahman MR, Islam K. Massive diphtheria outbreak among Rohingya refugees: lessons learnt. J Travel Med. 2019;26(1):1-3.
18. Dureab F, Müller O, Jahn A. Resurgence of diphtheria in Yemen due to population movement. J. Travel Med. 2018;25(1):1–2.
19. Zakikhany K, Efstratiou A. Diphtheria in Europe: current problems and new challenges. Future Microbiol. 2012;7:595–607.
20. Meinel DM, et al. Outbreak investigation for toxigenic Corynebacterium diphtheriae wound infections in refugees from Northeast Africa and Syria in Switzerland and Germany by whole genome sequencing. Clin. Microbiol. Infect. Off. Publ. Eur. Soc. Clin. Microbiol. Infect. Dis. 2016;22:1003.e1–1003.e8.
21. Scheifer C, et al. Re-emergence of Corynebacterium diphtheriae. Med Mal Infect. 2019;49:463–6.
22. Möller J, et al. Proteomics of diphtheria toxoid vaccines reveals multiple proteins that are immunogenic and may contribute to protection of humans against Corynebacterium diphtheriae. Vaccine. 2019;37:3061–70.
23. Zasada, A. A. Antimicrobial susceptibility and treatment. in *Corynebacterium diphtheriae* and related toxigenic species genomics, pathogenicity and applications. Burkovski A., editor. 239–246 (New York: Springer, 2014).
24. Kupferschmidt K. Life-saving diphtheria drug is running out. Science. 2017;355:118–9.
25. PAHO/WHO, P. A. H. O. / W. H. O. Epidemiological update: diphtheria. (2018).
26. Kneen R, et al. Penicillin vs erythromycin in the treatment of diphtheria. Clin Infect Dis Off Publ Infect Dis Soc Am. 1998;27:845–50.
27. Pereira GA, et al. Antimicrobial resistance among Brazilian Corynebacterium diphtheriae strains. Mem Inst Oswaldo Cruz. 2008;103:507–10.
28. von Hunolstein C, Scopetti F, Efstratiou A, Engler K. Penicillin tolerance amongst non-toxigenic Corynebacterium diphtheriae isolated from cases of pharyngitis. J Antimicrob Chemother. 2002;50:125–8.
29. Benamrouche N, et al. Microbiological and molecular characterization of Corynebacterium diphtheriae isolated in Algeria between 1992 and 2015. Clin. Microbiol. Infect. Off. Publ. Eur. Soc. Clin. Microbiol. Infect. Dis. 2016;22:1005.e1–1005.e7.
30. Paveenkittiporn W, Sripakdee S, Koobkratok O, Sangkitporn S, Kerdsin A. Molecular epidemiology and antimicrobial susceptibility of outbreak-associated Corynebacterium diphtheriae in Thailand, 2012. Infect. Genet Evol J Mol Epidemiol Evol Genet Infect Dis. 2019;75:104007.
31. Husada D, et al. First-line antibiotic susceptibility pattern of toxigenic Corynebacterium diphtheriae in Indonesia. BMC Infect Dis. 2019;19:1049.
32. Bernard K, Pacheco AL. In vitro activity of 22 antimicrobial agents against Corynebacterium and microbacterium species referred to the Canadian National Microbiology Laboratory. Clin Microbiol Newsl. 2015;37:187–98.
33. Schiller J, Groman N, Coyle M. Plasmids in Corynebacterium diphtheriae and diphtheroids mediating erythromycin resistance. Antimicrob Agents Chemother. 1980;18:814–21.
34. Jellard CH, Lipinski AE. Corynebacterium diphtheriae resistant to erythromycin and lincomycin. Lancet Lond Engl. 1973;1:156.
35. Patey O, et al. Antibiotic susceptibilities of 38 non-toxigenic strains of Corynebacterium diphtheriae. J Antimicrob Chemother. 1995;36:1108–10.
36. Tauch A, Bischoff N, Brune I, Kalinowski J. Insights into the genetic organization of the Corynebacterium diphtheriae erythromycin resistance plasmid pNG2 deduced from its complete nucleotide sequence. Plasmid. 2003;49:63–74.
37. Barraud O, Badell E, Denis F, Guiso N, Ploy M-C. Antimicrobial drug resistance in Corynebacterium diphtheriae mitis. Emerg Infect Dis. 2011;17:2078–80.
38. Cerdeño-Tárraga AM, et al. The complete genome sequence and analysis of Corynebacterium diphtheriae NCTC13129. Nucleic Acids Res. 2003;31:6516–23.
39. Farfour E, et al. Characterization and comparison of invasive Corynebacterium diphtheriae isolates from France and Poland. J Clin Microbiol. 2012;50:173–5.
40. Grimont PAD, et al. International nomenclature for Corynebacterium diphtheriae ribotypes. Res Microbiol. 2004;155:162–6.
41. Park WH, Williams AW. The production of diphtheria toxin. J Exp Med. 1896;1:164–85.
42. Dazas M, Badell E, Carmi-Leroy A, Criscuolo A, Brisse S. Taxonomic status of Corynebacterium diphtheriae biovar Belfanti and proposal of Corynebacterium belfantii sp. nov. Int J Syst Evol Microbiol. 2018;68:3826–31.
43. Badell, E. *et al.* Corynebacterium rouxii sp. nov., a novel member of the diphtheriae species complex. Res. Microbiol. 2020. https://doi.org/10.1016/j.resmic.2020.02.003.
44. Hauser D, Popoff MR, Kiredjian M, Boquet P, Bimet F. Polymerase chain reaction assay for diagnosis of potentially toxinogenic Corynebacterium diphtheriae strains: correlation with ADP-ribosylation activity assay. J Clin Microbiol. 1993;31:2720–3.
45. Engler KH, Glushkevich T, Mazurova IK, George RC, Efstratiou A. A modified Elek test for detection of toxigenic corynebacteria in the diagnostic laboratory. J Clin Microbiol. 1997;35:495–8.
46. Magiorakos AP, et al. Multidrug-resistant, extensively drug-resistant and pandrug-resistant bacteria: an international expert proposal for interim

standard definitions for acquired resistance. Clin Microbiol Infect. 2012;18: 268–81.

47. Badell E, et al. Improved quadruplex real-time PCR assay for the diagnosis of diphtheria. J Med Microbiol. 2019;68:1455–65.
48. Criscuolo, A. & Brisse, S. AlienTrimmer: a tool to quickly and accurately trim off multiple short contaminant sequences from high-throughput sequencing reads. Genomics. 2013. https://doi.org/10.1016/j.ygeno.2013.07.011.
49. Crusoe MR, et al. The khmer software package: enabling efficient nucleotide sequence analysis. F1000Research. 2015;4:900.
50. Liu Y, Schröder J, Schmidt B. Musket: a multistage k-mer spectrum-based error corrector for Illumina sequence data. Bioinforma Oxf Engl. 2013;29: 308–15.
51. Bankevich A, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19:455–77.
52. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinforma. Oxf. Engl. 2014;30:2068–9.
53. Page AJ, et al. Roary: rapid large-scale prokaryote pan genome analysis. Bioinforma. Oxf. Engl. 2015;31:3691–3.
54. Didelot X, Wilson DJ. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. PLoS Comput Biol. 2015;11:e1004041.
55. Guindon S, et al. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 2010;59:307–21.
56. Minh BQ, et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol. 2020;37:1530–4.
57. Collins C, Didelot X. A phylogenetic method to perform genome-wide association studies in microbes that accounts for population structure and recombination. PLoS Comput Biol. 2018;14:e1005958.
58. Sogues A, et al. Essential dynamic interdependence of FtsZ and SepF for Z-ring and septum formation in Corynebacterium glutamicum. Nat Commun. 2020;11:1641.
59. Mitchell AL, et al. InterPro in 2019: improving coverage, classification and access to protein sequence annotations. Nucleic Acids Res. 2019;47:D351–60.
60. Madeira F, et al. The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic Acids Res. 2019;47:W636–41.
61. Ansari MA, Didelot X. Inference of the properties of the recombination process from whole bacterial genomes. Genetics. 2014;196:253–65.
62. Santos AS, et al. Searching whole genome sequences for biochemical identification features of emerging and reemerging pathogenic Corynebacterium species. Funct Integr Genomics. 2018;18:593–610.
63. Tagini F, et al. Distinct genomic features characterize two clades of Corynebacterium diphtheriae: proposal of Corynebacterium diphtheriae subsp. diphtheriae subsp. nov. and Corynebacterium diphtheriae subsp. lausannense subsp. nov. Front. Microbiol. 2018;9:1743.
64. Kahlmeter G, et al. European harmonization of MIC breakpoints for antimicrobial susceptibility testing of bacteria. J Antimicrob Chemother. 2003;52:145–8.
65. Tauch A, Kassing F, Kalinowski J, Pühler A. The Corynebacterium xerosis composite transposon Tn5432 consists of two identical insertion sequences, designated IS1249, flanking the erythromycin resistance gene ermCX. Plasmid. 1995;34:119–31.
66. Lavollay M, et al. The beta-lactam-sensitive D,D-carboxypeptidase activity of Pbp4 controls the L,D and D,D transpeptidation pathways in Corynebacterium jeikeium. Mol. Microbiol. 2009;74:650–61.
67. Seth-Smith HMB, Egli A. Whole genome sequencing for surveillance of diphtheria in low incidence settings. Front Public Health. 2019;7:235.
68. Barksdale L. Corynebacterium diphtheriae and its relatives. Bacteriol Rev. 1970;34:378–422.
69. McLeod, J. W. The types mitis, intermedius and gravis of corynebacterium diphtheriae: a review of observations during the past ten years. Bacteriol Rev 7, 1–41 (1943).
70. Zou J, et al. Phenotypic and genotypic correlates of penicillin susceptibility in nontoxigenic Corynebacterium diphtheriae, British Columbia, Canada, 2015-2018. Emerg Infect Dis. 2020;26:97–103.
71. Marosevic DV, et al. Antimicrobial susceptibility of Corynebacterium diphtheriae and Corynebacterium ulcerans in Germany 2011-17. J Antimicrob Chemother. 2020. https://doi.org/10.1093/jac/dkaa280.
72. Sariadji K, Puspandari SN, Sembiring M. Antibiotic susceptibility pattern of Corynebacterium diphtheriae isolated from outbreaks in Indonesia 2010-2015. Indones Biomed J. 2018;10:51–5.
73. Parande MV, Roy S, Mantur BG, Parande AM, Shinde RS. Resurgence of diphtheria in rural areas of North Karnataka, India. Indian J Med Microbiol. 2017;35:247–51.
74. Anonymous. Conduite à tenir lors de l’apparition d’un cas de diphtérie. https://www.hcsp.fr/explore.cgi/hcspr20110304_conduitediphterie.pdf. (2011).
75. Tauch A, Götker S, Pühler A, Kalinowski J, Thierbach G. The 27.8-kb R-plasmid pTET3 from Corynebacterium glutamicum encodes the aminoglycoside adenyltransferase gene cassette aadA9 and the regulated tetracycline efflux system Tet 33 flanked by active copies of the widespread insertion sequence IS6100. Plasmid. 2002;48:117–29.
76. Tauch A, Kassing F, Kalinowski J, Pühler A. The erythromycin resistance gene of the Corynebacterium xerosis R-plasmid pTP10 also carrying chloramphenicol, kanamycin, and tetracycline resistances is capable of transposition in Corynebacterium glutamicum. Plasmid. 1995;33:168–79.
77. Yoon S, Kim H, Lee Y, Kim S. Bacteremia caused by Corynebacterium amycolatum with a novel mutation in gyrA gene that confers high-level quinolone resistance. Korean J Lab Med. 2011;31:47–8.
78. Sierra JM, Martinez-Martinez L, Vázquez F, Giralt E, Vila J. Relationship between mutations in the gyrA gene and quinolone resistance in clinical isolates of Corynebacterium striatum and Corynebacterium amycolatum. Antimicrob Agents Chemother. 2005;49:1714–9.
79. Pivot D, et al. Carriage of a single strain of non-toxigenic Corynebacterium diphtheriae biovar Belfanti (Corynebacterium belfantii) in four patients with cystic fibrosis. J Clin Microbiol. 2019. https://doi.org/10.1128/JCM.00042-19.
80. Forde BM, et al. Fatal respiratory diphtheria caused by β-lactam-resistant Corynebacterium diphtheriae. Clin Infect Dis Off Publ Infect Dis Soc Am. 2020. https://doi.org/10.1093/cid/ciaa1147.
81. Fishovitz J, et al. Disruption of allosteric response as an unprecedented mechanism of resistance to antibiotics. J Am Chem Soc. 2014;136:9814–7.
82. Otero LH, et al. How allosteric control of Staphylococcus aureus penicillin binding protein 2a enables methicillin resistance and physiological function. Proc Natl Acad Sci U S A. 2013;110:16808–13.
83. Bernardo-García N, et al. Allostery, recognition of nascent peptidoglycan, and cross-linking of the cell wall by the essential penicillin-binding protein 2x of Streptococcus pneumoniae. ACS Chem Biol. 2018;13:694–702.
84. Chewapreecha C, et al. Comprehensive identification of single nucleotide polymorphisms associated with beta-lactam resistance within pneumococcal mosaic genes. PLoS Genet. 2014;10:e1004547.
85. Brisse, S., Badell, E., Hennart, M. & Criscuolo, A. Genomic content and phylogenetic diversity of C. diphtheriae and related species were compared to clarify taxonomic relationships and study evolutionary trends. European Nucleotide Archive https://www.ebi.ac.uk/ena/browser/view/PRJEB22103 (2020).

Publisher’s Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

BMC

View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Directory of Open Access Journals

frontiers
in Pharmacology

MINI REVIEW
published: 21 June 2021
doi: 10.3389/fphar.2021.684162

Pharmacogenomics of NSAID-Induced Upper Gastrointestinal Toxicity

L. McEvoy, D. F. Carr* and M. Pirmohamed

Department of Pharmacology and Therapeutics, University of Liverpool, Liverpool, United Kingdom

Non-steroidal anti-inflammatory drugs (NSAIDs) are a group of drugs which are widely used globally for the treatment of pain and inflammation, and in the case of aspirin, for secondary prevention of cardiovascular disease. Chronic non-steroidal anti-inflammatory drug use is associated with potentially serious upper gastrointestinal adverse drug reactions (ADRs) including peptic ulcer disease and gastrointestinal bleeding. A few clinical and genetic predisposing factors have been identified; however, genetic data are contradictory. Further research is needed to identify clinically relevant genetic and non-genetic markers predisposing to NSAID-induced peptic ulceration.

Keywords: pharmacogenomic, NSAID (non-steroidal anti-inflammatory drug), gastrointestinal toxicities, aspirin, adverse drug reaction

OPEN ACCESS

Edited by:
*José A G Agúndez,
University of Extremadura, Spain*

Reviewed by:
*Ioannis S Vizirianakis,
Aristotle University of Thessaloniki,
Greece
Volker Martin Lauschke,
Karolinska Institutet (KI), Sweden*

***Correspondence:**
*D. F. Carr
D.Carr@liverpool.ac.uk*

Specialty section:
*This article was submitted to
Pharmacogenetics and
Pharmacogenomics,
a section of the journal
Frontiers in Pharmacology*

Received: *22 March 2021*
Accepted: *11 May 2021*
Published: *21 June 2021*

Citation:
*McEvoy L, Carr DF and Pirmohamed M
(2021) Pharmacogenomics of NSAID-
Induced Upper
Gastrointestinal Toxicity.
Front. Pharmacol. 12:684162.
doi: 10.3389/fphar.2021.684162*

INTRODUCTION

NSAIDs comprise a heterogeneous group of non-opioid drugs with effective analgesic, anti-inflammatory and antipyretic properties (Calatayud and Esplugues, 2016). They are employed in the treatment of acute and chronic pain conditions characterized by inflammation. While aspirin was used as an analgesic in the past, its main use nowadays is as an antithrombotic particularly at low doses, and for cancer prevention (Shaheen et al., 2002; Sandler et al., 2003; Bashir et al., 2019). Overall, NSAIDs are well tolerated, especially when used short-term (Ong et al., 2007), but because of the enormous usage globally, they are often implicated in adverse drug reactions.

NSAIDs, including low dose aspirin (LDA), are one of the most commonly prescribed classes of medication, accounting for approximately 5–10% of prescriptions globally (Onder et al., 2004). Two decades ago, >30 million people were estimated to take NSAIDs daily (Singh and Triadafilopoulos, 1999). Pharmacoepidemiological studies indicate that NSAID use is increasing. A 2010 National Health Interview Survey (CDC, United States) reported increases of 57 and 41% in aspirin and NSAID use, respectively, over 5 years (Zhou et al., 2014), but this may still be an underestimate given the wide availability of aspirin and ibuprofen as over-the-counter (OTC) formulations. Telephone surveys of United States OTC NSAID users found that drugs were often used/taken inappropriately with 26% of respondents exceeding recommended doses (Goldstein and Lefkowith, 1998; Wilcox

---

Abbreviations: AA, Arachidonic acid; ADR, Adverse drug reaction; ASA, (Acetylsalicylic acid) Aspirin; COX, cyclo-oxygenase; Coxibs, COX-2 inhibitors; CPIC, Clinical Pharmacogenetics Implementation Consortium; CV, Cardiovascular; CYP, Cytochrome P450; FMT, Fecal microbiota transplant; GI, Gastrointestinal; GWAS, Genome wide association study; IBD, Inflammatory bowel disease; LD, Linkage disequilibrium; LDA, Low dose aspirin; NSAID, Non-steroidal anti-inflammatory drug; OTC, Over-the-counter; PG, Prostaglandin; PGx, Pharmacogenomics; PhV, Pharmacovigilance; PUD, Peptic ulcer disease; SNP, Single nucleotide polymorphism; TX, Thromboxane; tNSAID, Traditional NSAID; UC, Ulcerative colitis; UGIB, Upper gastrointestinal bleeding; UGT, Uridine 5′-diphosphate glucuronosyltransferases (UDP-glucuronosyltransferase).

et al., 2005; Goldstein and Cryer, 2015). In Germany, analgesic use significantly increased from 19.2% in 1998 to 21.4% in 2008–2011. This rise was found to be attributed exclusively to the use of OTC formulations increasing from 10.0 to 12.2% (prescribed analgesic use remained constant at 7.9%). Ibuprofen was most commonly used, followed by aspirin (Sarganas and et al., 2015). Higher frequencies of ibuprofen use have also been documented in Denmark (Olsen and et al., 2011) and Spain (Gómez-Acebo and et al., 2018). In contrast, however, diclofenac was reported as the most frequently used NSAID (followed by ibuprofen) in a study across 15 countries (Australia, Bangladesh, Canada, China, China (Hong Kong), England, Indonesia, Malaysia, New Zealand, Pakistan, Philippines, Singapore, Taiwan, Thailand, and Vietnam) (McGettigan and Henry, 2013). The trend for increasing OTC analgesic use has also been echoed in Australia. Between 2001 and 2009, there was a 15% increase in the use of ibuprofen, naproxen and diclofenac (Stosic et al., 2011). In general practice, NSAID use by the elderly is prolific, reported at 96% (96% in males, 96.7% in females) in patients >65 years (Pilotto et al., 2003).

NSAIDs are responsible for ~30% of ADR hospitalisations (Pirmohamed et al., 2004); cardiovascular, gastrointestinal (GI) and renal complications are associated with their use (Bhala et al., 2013; Szeto et al., 2020). Estimates suggest that 5,000–16,500 deaths in the United States and 400–1,000 deaths in the United Kingdom are a direct consequence of NSAID-induced upper GI ulceration and bleeding annually (Wolfe et al., 1999; Langman, 2001; Hawkey and Langman, 2003).

Genetic factors predisposing to NSAID-induced upper GI toxicity have been described, yet findings have been inconsistent and contradictory. This mini review discusses current literature and seeks to identify areas to focus collaborative efforts in the field.

NSAID-INDUCED UPPER GASTROINTESTINAL TOXICITY

The first serious NSAID-induced adverse event to be identified was upper GI injury (Walt et al., 1986; Langman, 1988). It is also recognized as one of the most predominant ADRs in the United States (Butt et al., 1988; Fries, 1996). NSAIDs and aspirin have now overtaken *Helicobacter pylori* as the principal cause of GI toxicity in western countries (Musumba et al., 2009; Bjarnason, 2013). "Gastrointestinal toxicity" refers to a collection of heterogeneous pathologies affecting various tissues and organs of the GI tract.

Clinical Problem and Disease Burden

NSAID-related ADRs range from mild to severe and can result in death. Cryer (2004) conveniently stratified typical manifestations of NSAID-induced GI injury in to three tiers: dyspepsia; asymptomatic ulceration; and more serious complications (GI bleeding, perforation, obstruction, symptomatic ulceration) (Cryer, 2004). Although upper GI events are more common, the entirety of the GI tract may be affected (Maiden et al., 2005). Lower GI complications are not as well defined (Langman et al., 1985). In the large intestine, non-specific colitis, increased gut permeability, malabsorption and bleeding have been reported. Inflammatory bowel disease (IBD) (Bjorkman, 1998; Faucheron, 1999; Forrest et al., 1999) and diverticular disease (Campbell and Steele, 1991) may also be exacerbated by NSAIDs. In this article, we only focus on upper GI events.

Patient characteristics increasing the risk of NSAID-induced GI events have been identified (summarized in **Table 1**). These include: advanced age (Fries et al., 1991; Hernández-Díaz and Rodríguez, 2000; Russell, 2001; Sostres et al., 2013; Chi et al., 2018); *H. pylori* infection (Leontiadis and et al., 2007; Sostres et al., 2010); multimorbidity/comorbidity (Chi et al., 2018; Jankovic et al., 2009; Weil and et al., 2000; Kim, 2015); polypharmacy (Davies and Wallace, 1996) and concomitant medications (de Abajo et al., 1999; Silverstein et al., 2000; Sorensen et al., 2000; Garcia Rodriguez and Hernández-Díaz, 2001; Johnsen et al., 2001; Lazzaroni and Bianchi Porro, 2001; Laine et al., 2002; de Jong et al., 2003; Helin-Salmivaara et al., 2007; Lanas et al., 2007; Åhsberg et al., 2010; Masclee et al., 2013; Sostres et al., 2013; Anglin et al., 2014; Olsen et al., 2020). Studies have shown that the risk of NSAID-induced GI complications is dose-dependent (Silverstein and et al., 1995; Bombardier et al., 2000; Laporte et al., 2004; González-Pérez and et al., 2014; Figueiras and et al., 2016) and remains linear over time (Silverstein and et al., 1995; Bombardier et al., 2000; Rostom et al., 2007; Goldstein et al., 2011).

Increased/prolonged exposure (habitual in chronic pain conditions through high-dose and multiple-NSAID use) elevates the risk and/or severity of toxicity (Garcia Rodriguez and Hernández-Díaz, 2001; Bhala et al., 2013; Lanas et al., 2015). Treatment guidelines recommend the lowest effective dose for the shortest period of time. Whilst it is recommended that long-term NSAID use should be avoided (Bhatt et al., 2008; Lanza et al., 2009), this is difficult in practice for many patients. Individuals with a past history of GI injury (including uncomplicated or complicated ulcers) are considered at the highest risk of complications (Sostres et al., 2013; Chi et al., 2018; Silverstein and et al., 1995; Van Hecken et al., 2000; Laine, 2001; Laine, 2006). COX-2 inhibitors (Coxibs) and gastroprotective therapies such as proton pump inhibitors (effective anti-secretory agents) help to mitigate the risk of GI injury (Gargallo et al., 2014; Melcarne et al., 2016).

Mild upper GI symptoms are reported by up to 40% of NSAID users (Larkai et al., 1989; Hirschowitz, 1994). The most clinically significant upper GI NSAID-induced ADRs are symptomatic and/or complicated peptic ulcer. Symptomatic peptic ulcer incidence rates in chronic NSAID users are between 2–4% annually (1–3% for serious ulcer/upper GI complications) (Cryer, 2004; Silverstein et al., 2000; Silverstein and et al., 1995; Bombardier et al., 2000; Blower and et al., 1997; Paulus, 1988), a threefold to fivefold increase compared to non-users (García Rodríguez and Jick, 1994; Gutthann et al., 1997; Hernández-Díaz and Rodríguez, 2000; Cryer, 2004; Gevers et al., 2014).

It is important to note that symptoms do not necessarily correlate with the severity of peptic ulcer disease and/or its complications (Sostres et al., 2013)—50–60% of patients can

TABLE 1 | Reported risk factors for developing NSAID-induced upper GI toxicity.

Patient characteristic	Description	References
Patient factors		
Advanced age	>60 years	Fries. (1996), Hernández-Díaz and Rodríguez, (2000), Chi et al. (2018)
	>70 years	Sostres et al. (2013), Russell. (2001)
Comorbidity/ multimorbidity	Renal failure (receiving haemodialysis)	Jankovic et al. (2009)
	Helicobacter pylori infection	Leontiadis et al. (2007), Sostres et al. (2010)
	Diabetes mellitus	Weil and et al. (2000), Kim (2015)
	Cardiovascular disease	Chi et al. (2018)
Clinical history	Previous upper GI clinical event	Silverstein et al. (1995), Laine (2001), Laine (2006), Van Hecken et al. (2000), Sostres et al. (2013), Chi et al. (2018)
Drug factors		
Increased NSAID exposure	High dose/prolonged exposure NSAID therapy	Silverstein et al. (1995), Bombardier et al. (2000), Laporte et al. (2004) Garcia Rodriguez and Hernández-Díaz, (2001), Bhala et al. (2013), Lanas et al. (2015)
Polypharmacy	Polypharmacy	Davies and Wallace, (1996)
Concomitant drug use	Aspirin	Sorensen et al. (2000), Silverstein et al. (2000), Garcia Rodriguez and Hernández-Díaz, (2001), Åhsberg et al. (2010), Lazzaroni and Bianchi Porro, (2001)
	Non-aspirin antiplatelets	Sostres et al. (2013)
	Anticoagulants	Johnsen et al. (2001), Lanas et al. (2007), Olsen et al. (2020)
	Oral corticosteroids	Garcia Rodriguez and Hernández-Díaz, (2001), Laine et al. (2002), Masclee et al. (2013)
	Selective serotonin reuptake inhibitors	de Abajo et al. (1999), de Jong et al. (2003), Helin-Salmivaara et al. (2007), Anglin et al. (2014)

Abbreviations: GI, gastrointestinal; NSAID, non-steroidal anti-inflammatory drug.

be asymptomatic (including having unremarkable endoscopies (McCarthy, 1989)) prior to developing potentially fatal NSAID-induced complicated ulcers (Armstrong and Blower, 1987). A prospective observational study found this figure to be as high as 80% (Singh and et al., 1996). In NSAID users who do develop lesions (endoscopically confirmed subepithelial hemorrhages, erosions, and ulcers), healing often occurs before symptoms manifest—"silent ulceration"—because of efficient gastroprotective mechanisms (Larkai et al., 1987; Lanas and Hunt, 2006; Sostres et al., 2013; Carr et al., 2017). Notwithstanding, peptic ulcer prevalence ranges from 15–40% in chronic NSAID users (McCarthy, 1989; Geis et al., 1991; Dajani and Agrawal, 1995; McCarthy, 1998). Gastric ulcer risk is slightly higher than duodenal ulcer risk (McCarthy, 1989; Larkai et al., 1987; McCarthy, 1998; Henry et al., 1993; Gabriel et al., 1991; Griffin and et al., 1991). Endoscopy remains the gold standard for PUD diagnosis (Dunlap and Patterson, 2019).

LDA is extensively used as prophylactic therapy against cardiovascular thrombotic events in high-risk individuals (Collaboration., 2002; Patrono et al., 2004). Throughout Europe and the United States, approximately 23% of the population use aspirin at least once per week (Larson et al., 2005). A prospective analysis of ADRs requiring hospitalization found that aspirin was the most commonly implicated, responsible for 18% of admissions; 74% of these patients were taking LDA. GI bleeding was the most common adverse event, observed in 72% of aspirin-related admissions (Pirmohamed et al., 2004). This is most likely to be due to aspirin-associated peptic ulceration (Niv et al., 2005). Generally, individuals on LDA therapy are older and multimorbid, increasing the likelihood of polypharmacy; some patients take both aspirin and NSAID(s) (Murray and et al., 1998) which is known to further raise the risk of GI toxicity and bleeding (Silverstein et al., 2000; Sorensen et al., 2000; Garcia Rodriguez and Hernández-Díaz, 2001). GI symptoms in chronic LDA therapy are challenging and are a reason for discontinuation of treatment and non/poor compliance (Cayla et al., 2012). Adverse events have been cited as responsible for LDA-discontinuation in almost 50% of patients (Herlitz et al., 2010). Discontinuation is associated with increased risk of new CV events (Biondi-Zoccai et al., 2006; Sostres and Lanas, 2011) and should only be advocated after sound clinical assessment when the risk of GI bleeding outweighs that of CV events.

NSAIDs differ in their propensity to cause upper GI injury based on their COX-1/COX-2 selectivity. Due to selective COX-2 inhibition, coxibs offer greater GI safety and are associated with a lower risk of upper GI injury and clinically significant ulcer complications compared to traditional NSAIDs (tNSAIDs) (Rostom et al., 2007; Chan et al., 2010). A meta-analysis by Castellsague et al. (2012) calculated the relative risk (RR) of upper GI complications as follows:

- RR <2 aceclofenac, celecoxib, ibuprofen.
- RR 2 to <4 diclofenac, ketoprofen, meloxicam, nimesulide, rofecoxib, sulindac.
- RR 4—5 diflunisal, indomethacin, naproxen, tenoxicam.
- RR >5 azapropazone, ketorolac, piroxicam (Castellsague et al., 2012).

Celecoxib has been shown to be associated with a lower risk of upper-GI events (Silverstein et al., 2000; García-Rayado et al., 2018) whilst celecoxib plus proton pump inhibitor is preferred over tNSAID (naproxen) plus proton pump inhibitor to reduce the risk of recurrent upper GI bleeding in individuals requiring concomitant aspirin and NSAID (Chan et al., 2017).

Mechanisms of NSAID-Induced Gastrointestinal Toxicity

NSAID pharmacology, first described by Vane. (1971), is well documented (Vane, 1971). The cyclo-oxygenase (COX) pathway is responsible for the biosynthesis of prostanoids, including prostaglandins (PGs) and thromboxane (TX), from arachidonic acid (AA). Two clinically significant COX isoenzymes exist, COX-1 and COX-2, encoded by the genes *PTGS1* and *PTGS2,* respectively (Plaza-Serón et al., 2018). Prostanoids are pro-inflammatory mediators: PGs produced by COX-1 (and to a lesser degree, COX-2 (Wallace, 2008)) are essential in maintaining/regulating gastroprotection (Dickman and Ellershaw, 2004). Fundamentally, PG generation is forestalled by NSAIDs via reversible inhibition (except aspirin, an irreversible inhibitor (Funk et al., 1991; Catella-Lawson et al., 2001)) of COX isoenzymes, thereby decreasing inflammation.

NSAIDs are sub-divided based on COX-selectivity: tNSAIDs are non-selective, inhibiting both isoenzymes. COX-2 selective inhibitors (coxibs) are isoform-specific, designed to help mitigate the GI adverse events of tNSAIDs whilst retaining anti-inflammatory and analgesic activity. Inhibition of COX-2 is the desired therapeutic aim of NSAIDs. Upper GI toxicity is more common with tNSAIDs than with coxibs (Pilotto et al., 2005).

Under normal conditions, gut homeostasis is maintained. However, imbalance(s) in the gastroduodenal mucosal lining may result in GI injury. Protective factors comprise the gastric epithelial cells and hydrophobic mucus-bicarbonate bilayer, cellular repair, remodeling, restitution, and adequate blood supply–all of these are PG-regulated. Microvascular damage reduces gastric mucosal blood flow, an initial and crucial event in ulcer pathogenesis (Musumba et al., 2009; Bjarnason et al., 2018). The locale of greatest localized ischemia correlates with the site of most NSAID-induced ulceration, the gastric antrum (Musumba et al., 2009). Gastroprotective strategies can mitigate the risk of PUD, but will not be discussed further in this review: Wallace. (2008), Musumba et al. (2009), Bjarnason et al. (2018) provide excellent analyses (Wallace, 2008; Musumba et al., 2009; Bjarnason et al., 2018).

The pathogenesis of NSAID-induced GI injury is complex and multifactorial and can be divided into two broad areas: 1) Topical mechanisms (direct damage to GI mucosa). 2) Systemic mechanisms (COX-inhibition mediated).

1) *Topical effects*—NSAIDs can cause disturbances in the gastric mucosal epithelium initially through the development of erosions. NSAID toxicity is dependent on their physiochemical properties as lipid-soluble weak organic acids (pKa 3–5) (Brune et al., 1976; Brune and Graf, 1978; Rainsford and Whitehouse, 1980; Musumba et al., 2009; Bjarnason et al., 2018). Detergent properties allow interaction with surface membrane phospholipids, disrupting the mucosal barrier and provoking superficial injury (Lichtenberger, 2001). This permits NSAID to move from the lumen (low gastric pH measuring 1.5–3.5) into epithelial cells (pH neutral, 6.5–7.0), kickstarting disruption of the cellular metabolic pathways culminating in dysfunction, cytotoxic events and apoptotic pathway activation (Musumba et al., 2009; Handa et al., 2014).
2) *Systemic effects*—Robust gastric mucosal defence/repair is heavily dependent upon perpetual synthesis of COX-derived prostanoids. COX-1 is expressed in most tissues performing "housekeeping duties," maintaining and regulating various physiological functions. In the GI tract, COX-1 is abundant, producing PGs and TXs involved in cytoprotection. COX-2 is expressed at low levels in the intact stomach, but rapidly induced by COX-1 inhibition or inflammatory stimuli/injury, and produced in vast quantities by cytokines and hormones (Rainsford, 2007; Ricciotti and FitzGerald, 2011; Wongrakpanich et al., 2018).

Classically, GI complications were primarily attributed to COX inhibition: decreased PG levels resulting in increased gastric acid secretion, suppressed mucus and bicarbonate secretion, decreased mucosal blood flow and decreased cell proliferation (Cohen, 1987; Wallace, 1992; Wallace, 2008) resulting in compromised mucosal defense and delayed healing. Selective COX-1 inhibition leads to suppression of mucosal PG production by 95–98% without observable inflammation or ulceration (Ligumsky et al., 1983; Ligumsky et al., 1990; Langenbach et al., 1995; Sigthorsson et al., 2002). Inhibition of COX-2 has yielded similar results (Morham et al., 1995; Sigthorsson et al., 2002). However dual inhibition of both isoenzymes by NSAIDs induces severe gastric lesions (Wallace et al., 2000; Gretzer et al., 2001; Tanaka and et al., 2001). Both COX isoenzymes are also important in ulcer healing (Tanaka et al., 2002; Schmassmann et al., 2006; To et al., 2001; Bhandari et al., 2005; Starodub et al., 2008; Chan and et al., 2005).

In isolation, topical mechanisms are unlikely to induce notable GI toxicity. In combination with systemic factors, however, NSAIDs provoke an imbalance that incites mucosal injury. Cellular damage translates into tissue damage. COX-1 inhibition dramatically suppresses gastroprotection leading to microvascular damage, ischemia, and decreased mucus production. A complex interplay between factors ignites a cascade, involving many mechanisms that disrupt gastroprotection, engendering ulcerogenic conditions.

GENETIC RISK FACTORS

It is well acknowledged that interindividual variability exists in response to drugs, including NSAIDs, which may at least be genetic in origin. Using conventional dosing regimens in analgesia, for example, some individuals will have inadequate pain control while others will encounter toxicity from the same dose (Kapur et al., 2014). Genetic variants affecting treatment outcomes can be categorized into two broad types: 1) genes affecting drug pharmacokinetics (PK), and 2) genes affecting drug pharmacodynamics (PD) (Stamer et al., 2010; Kapur et al., 2014). It should however be noted that given the different mechanisms of action and metabolic pathways, the PK and PD genetic variability cannot be generalized across all NSAID classes. **Table 2** summarizes notable genetic risk factors for NSAID-induced upper GI toxicity.

Pharmacokinetic-Related Associations

Cytochrome P450 (CYP) gene polymorphisms are strongly associated with adverse drug reactions in general (Johansson

TABLE 2 | Notable genetic associations in NSAID-induced upper GI toxicity.

Genetic loci	Therapy	Phenotype	Association	_p_-value	Or (95%CI)	References
CYP2C9	NSAIDs—various	Endoscopically confirmed UGIB	_CYP2C9*2_ genotype (in heterozygosity or homozygosity) increases risk of GI bleeding	_p_ = 0.009	Crude OR = 1.92 (95% CI = 1.14–3.25)	Martínez et al. (2004)
CYP2C9	CYP2C9-metabolised NSAIDs	Endoscopically confirmed UGIB	Using _CYP2C9*1/*1_ wild type as control, significantly higher frequencies of bleeding observed in			Pilotto et al. (2007)
			CYP2C9*1/*3	_p_ = < 0.001	OR = 12.9 (95% CI = 2.917–57.922)	
			CYP2C9*1/*2	_p_ = 0.036	OR = 3.8 (95% CI = 1.090–13.190)	
			CYP2C9*3 allele carriers have significant risk of bleeding		Adjusted OR = 7.3 (95% CI = 2.058–26.004)	
CYP2C9	Non-aspirin NSAIDs	Endoscopically confirmed ulcers/bleeding erosions	_CYP2C9*3_ loss-of-function allele associated with acute UGIB in NSAIDs other than aspirin	_p_ = 0.0002	OR = 7.2 (95% CI = 2.6–20.3)	Carbonell et al. (2010)
CYP2C9	CYP2C9-metabolised NSAIDs	Endoscopically confirmed erosions/lesions/UGIB	_CYP2C9*3_ variant increases risk of UGIB for defined daily doses >0.5		_CYP2C9*3_ OR = 18.07 (95% CI = 6.34–51.53) Adjusted OR	Figueiras et al. (2016)
CYP2C8 and _CYP2C9_	CYP2C8/2C9-metabolised NSAIDs	Endoscopically confirmed UGIB	_CYP2C8*3_ and _CYP2C9*2_ exist in LD			Blanco et al. (2008)
			Combined _CYP2C8*3_ + _CYP2C9*2_ genotype associated with increased risk of bleeding	_p_ = 0.003	OR = 3.73 (95% CI = 1.57–8.88)	
CYP2C19	NSAIDs—various	Endoscopically confirmed UGIB	_CYP2C19*17_ associated with PUD, but not UGIB PUD distribution varied according to _CYP2C19*17_ genotype _*1/*1_, 64.3%; _*1/*17_, 71.7%; _*17/*17_, 73.8%	_p_ = 0.024	OR (additive model) = 1.47 (95% CI = 1.12–1.92)	Musumba et al. (2013)
COX-1	Aspirin and ethanol pre-treatment		SNPs _A-842G_ and _C50T_ in complete LD			Halushka et al. (2003)
			Heterozygous _A-842G/C50T_ haplotype shows significantly greater PG H (2) inhibition	_p_ = 0.01		
COX-1	Not stated	Endoscopically confirmed bleeding GU or DU	_A-842/C50T_ polymorphism has lower (yet non-significant) Risk of PU bleeding compared to wild type		OR = 0.5 (95% CI = 0.18–1.34) Adjusted* OR = 0.75 (95% CI = 0.19–3.01) Adjusted for: sex, age, smoking, NSAID/aspirin use, _H. pylori_ infection	van Oijen et al. (2006)
COX-2	Aspirin (ASA)	Surgical or endoscopic UGIB diagnosis	rs689466 T > C increases risk magnitude of UGIB in ASA users			Mallah et al. (2020)
			Variant carriers taking ASA v wild-type carriers NOT taking ASA	_p_ = 0.0022	OR = 8.22 (95% CI = 2.14–31.59)	
			Variant carriers taking ASA v wild-type carriers taking ASA	_p_ = 0.3036	OR = 2.98 (95% CI = 0.37–23.96) OR adjusted as per Mallah et al. (2020)	

Abbreviations: *ASA, (acetylsalicylic acid) aspirin; CI, confidence interval; DU, duodenal ulcer; GI, gastrointestinal; GU, gastric ulcer; LD, linkage disequilibrium; NSAID, non-steroidal anti-inflammatory drug; OR, odds ratio; PG H(2), prostaglandin H(2); PU, peptic ulcer; PUD, peptic ulcer disease; UGIB, upper gastrointestinal bleeding; SNP, single nucleotide polymorphism.*

and Ingelman-Sundberg, 2011; Sim et al., 2013). Collectively, CYP isoenzymes mediate ~80% of phase I metabolism of clinically relevant drugs (Evans and Relling, 1999; Eichelbaum et al., 2006; Petrović et al., 2020), including NSAIDs (Blanco et al., 2005). Phase I oxidative metabolism precedes phase II conjugative metabolism (glucuronidation) driven by uridine 5′-diphosphate glucuronosyltransferases (UDP-glucuronosyltransferase, UGTs) (Kuehl et al., 2005; Monrad et al., 2014; Sandson, 2015), supplemented by sulfate conjugation (sulfotransferases) (Ulrich et al., 2006; Loetsch and Oertel, 2013). NSAID metabolism varies and this can be due to CYP polymorphisms, which have been implicated in NSAID-induced ADRs including PUD and UGIB. Studies have also demonstrated that GI damage is dose-dependent (Laporte et al., 2004; Figueiras and et al., 2016).

NSAID metabolism is specifically associated with the CYP2C subfamily (Ingelman-Sundberg et al., 2007; Musumba et al., 2013) (including the highly polymorphic CYP2C8, CYP2C9, and CYP2C19 isoforms), which make up 20% of hepatic CYP450 content and metabolize 25–30% of clinically used drugs. Polymorphisms in these genes may result in modified expression or functionality, correlating with altered metabolism and clearance which may affect bioavailability (Dai et al., 2001; Kirchheiner et al., 2002; Musumba et al., 2013; Krasniqi et al., 2016; Guengerich, 2020). Celecoxib, ibuprofen, lornoxicam and piroxicam are extensively (>90%) metabolized by CYP2C enzymes (Agúndez et al., 2009).

Estimates suggest that CYP2C9 is responsible for biotransformation and metabolic clearance in 15–20% of all phase I metabolized drugs (Goldstein and de Morais, 1994; Schwarz, 2003; Van Booven et al., 2010; Niinuma et al., 2014). CYP2C9 is involved as the main or secondary enzyme in the metabolism of most NSAIDs, including aceclofenac, aspirin, celecoxib, diclofenac, flurbiprofen, indomethacin, lornoxicam, meloxicam, naproxen, piroxicam, and tenoxicam (Davies et al., 2000; Rodrigues, 2005; Agúndez et al., 2009; Samer et al., 2013; Theken et al., 2020). With celecoxib, lornoxicam and piroxicam, CYP2C9 is the predominant enzyme, responsible for 90% of drug metabolism (Agúndez et al., 2009).

At least 62 variant alleles and multiple sub-alleles have been reported for *CYP2C9* (PharmVar (https://www.pharmvar.org/gene/CYP2C9) (PharmVar, 2017; Gaedigk et al., 2018; Theken et al., 2020) which vary in frequency ethnically, geographically and racially (Theken et al., 2020; Theken et al., 2020; Gene-specific Information Tables for CYP2C9, 2020). The allelic variants *CYP2C9*2* (rs1799853) and *CYP2C9*3* (rs1057910) show high allele frequencies in human populations (García-Martín et al., 2006); the estimated prevalence in European populations is 14% for *CYP2C9*2* and 8% for *CYP2C9*3* (Lee et al., 2002; Xie et al., 2002; Sánchez-Diz et al., 2009). These variants, most notably *CYP2C9*3*, have been extensively studied and have been associated with decreased NSAID metabolism (Visser et al., 2005; García-Martín et al., 2006; Agúndez et al., 2009; Wadelius et al., 2009; Wang et al., 2011).

Carriers of low activity *CYP2C9* alleles could be at greater risk of GI toxicity due to increased NSAID exposure. Frustratingly, case-control studies investigating associations between *CYP2C9*2* and **3* variants and NSAID-induced ADRs have often been contradictory. Several studies have reported no associations between *CYP2C9* variants and NSAID-induced PUD and/or UGIB (Martin et al., 2001; Van Oijen et al., 2005; Vonkeman et al., 2006; Lopezrodriguez et al., 2008; Ma et al., 2008; Musumba et al., 2013; Ishihara et al., 2014), while others have reported that *CYP2C9* variants predispose to PUD (Martínez et al., 2004; Pilotto et al., 2007), including the presence of a gene-dose effect (Figueiras and et al., 2016), and that there may be a combined effect of *CYP2C8*/*CYP2C9* variation (Blanco et al., 2005; Blanco et al., 2008).

Efforts to provide PGx-based guidance on NSAIDs recently culminated in the Clinical Pharmacogenetics Implementation Consortium (CPIC) guideline for *CYP2C9* and NSAIDs (Theken et al., 2020). The guideline categorizes *CYP2C9* alleles as follows: Normal function (wild type) *CYP2C9*1*, decreased function *CYP2C9*2*, **5*, **8*, **11*, and no function *CYP2C9*3*, **6*, **13* (Theken et al., 2020; Theken et al., 2020; Gene-specific Information Tables for CYP2C9, 2020). *In vitro* and clinical studies have suggested that the "decreased-function" and "no-function" *CYP2C9* alleles are substrate-dependent. A meta-analysis of *CYPC9* variant alleles on NSAID exposure, based on predicted metaboliser phenotypes (poor metabolisers, intermediate metabolisers and normal metabolisers), showed that CYP2C9 poor metabolisers had decreased metabolic clearance, a prolonged plasma elimination half-life and increased plasma concentrations. Since NSAID-induced GI toxicity is known to be dose-dependent, the risk and severity of toxicity is likely to be increased (Bhala et al., 2013; Laporte et al., 2004; Figueiras and et al., 2016; Lanas et al., 2015). Hence, to mitigate the risk of PUD, dose-reduction, careful monitoring for toxicity/ADRs and alternative therapies, (e.g. non-CYP2C9-metabolized NSAIDs (such as naproxen)) should be used (Theken et al., 2020).

A subsequent systematic review and meta-analysis by Macías et al. (2020) also addressed discrepancies in previous studies. In an exhaustive study, they concluded: "There is a clear and consistent association of the development of GI adverse events with the *CYP2C9* genotype, and the association is slightly stronger in patients with GI bleeding." Consistent with this, Theken et al. (2020), found that the associations were stronger in poor metabolisers (again based on metaboliser phenotypes) with a significant gene-dose effect. Furthermore, allele-specific analyses showed that *CYP2C9*2* was a poor risk predictor (marginal effects) in contrast to *CYP2C9*3*, which clearly showed a highly significant association with increased risk of upper GI adverse events and GI bleeding (Macías et al., 2020). This supports the findings of Figueiras and et al. (2016) who demonstrated dose-dependency in NSAID-induced GI damage and the indication that the *CYP2C9*3* allele can be used as a predictive UGIB risk marker for CYP2C9-metabolised NSAIDS (Figueiras and et al., 2016).

CYP2C8 plays an accessory metabolic role for certain NSAIDs including ibuprofen and diclofenac (Agúndez et al., 2009; Garciamartin and et al., 2004). *CYP2C9*2* is in strong linkage disequilibrium (LD) with *CYP2C8*3* (Speed et al., 2009). Findings however have again been conflicting: no association between *CYP2C8*3* and UGIB was reported by Musumba *et al.* (Musumba et al., 2013) while others have reported a positive association with GI bleeding (Agúndez et al., 2009), including when combined with *CYP2C9*2* (Blanco et al., 2005; Blanco et al., 2008).

CYP2C19 plays a minor role in NSAID metabolism. Interestingly, we showed that the *CYP2C19*17* "gain of function" polymorphism was associated with PUD (Musumba et al., 2013), irrespective of the etiology of PUD. This may be related to the fact that CYP2C19 is involved in the metabolism of AA, and more extensive metabolism of AA may impair gastric mucosal defences. This is consistent with the fact that disrupted AA metabolism is involved in PUD pathogenesis (Musumba et al., 2009).

Thus, the role of the of the *CYP2C* gene locus in predisposing to PUD is complex and likely to vary with the NSAID given, its dose and the complement of SNPs the patient has at the *CYP2C* gene locus, including at *CYP2C9* (where low activity variants will increase exposure to NSAIDs) and the gain of function polymorphism at *CYP2C19* which increases the metabolism of gastro-protective AA.

Pharmacodynamic-Related Associations

COX enzymes produce PGs from AA, playing vital roles in gastric defence. COX enzymes are the primary target of NSAIDs. There is potential for peptic ulcer pathogenesis to be influenced by functional polymorphisms in the COX-encoding genes (Agúndez et al., 2015). Individuals carrying reduced function COX-enzymes may be potentially susceptible to NSAID-induced peptic ulceration (Musumba et al., 2009).

COX-2 single nucleotide polymorphism (SNPs) have been observed to affect responsiveness to celecoxib (Lee et al., 2017), but no association with NSAID-induced GI injury was reported. Studies of *COX-1* genetic polymorphisms and GI injury have also yielded contradictory findings (van Oijen et al., 2006; Arisawa et al., 2007). As such, there is insufficient evidence currently to determine whether the *COX* gene polymorphisms predispose to NSAID-induced PUD.

A recent preliminary study by Mallah et al. identified polymorphisms in platelet activity, angiogenesis and inflammatory response genes were associated with aspirin-related UGIB (Mallah and et al., 2020), building on the findings of previous investigations (Shiotani et al., 2010; Shiotani and et al., 2013; Shiotani et al., 2014; Wu et al., 2016; Cho and et al., 2016; Milanowski et al., 2017). The group identified “positive markers,” that indicated an increased risk of aspirin related UGIB and protective “negative markers” which decreased the risk (Mallah and et al., 2020). Of particular interest, rs689466 T > C, a SNP in the *COX-2* gene was associated with an increased risk of UGIB (Mallah and et al., 2020). However, these data are preliminary and need to be replicated in other cohorts.

There is significant literature describing associations between NSAID-induced immune-mediated ADRs and HLA alleles. These associations are with type I (immediate) urticarial and anaphylactic reactions (*HLA-DR11* and aspirin (Quiralte et al., 1999) and also type IV (delayed) reactions such as Stevens-Johnson syndrome/toxic epidermal necrolysis (*HLA-B*73:01* and oxicam (Lonjou et al., 2008)). To the best of our knowledge, there are no reports of HLA associations with NSAID-induced GI toxicity which is reflective of the non-immune mediated nature of the adverse event.

Based on the pharmacogenomic data presented herein, it is reasonable to suggest that evidence for genetic predisposition to NSAID-induced upper GI toxicity is somewhat contradictory. This reflects the complex nature of the etiology of this ADR but also the significant inter-individual variability in non-genetic risk factors. There are a number of factors which may influence inter-individual pre-disposition to NSAID GI and need to be integrated with pharmacogenomics to further our understanding of individual risk. These include, but not limited to: 1) the impact of altered GI physiology and subsequent drug disposition in special groups and the impact of disease physiology (Stillhart et al., 2020), 2) variability in physiological regulation via gastric enteroendocrine cells (Mace et al., 2015) and 3) *Helicobacter pylori* infection which is a significant risk factor for peptic ulceration independent of NSAID use (Huang et al., 2002).

BEYOND THE HUMAN GENOME

Spurred on by the completion of the Human Microbiome Project (2012), growing evidence supports the role of the microbiome in health and disease. Detrimental changes in gut microbiota composition (intestinal microbiota dysbiosis) are suggested as markers of GI pathogenesis in Crohn’s disease (Gevers et al., 2014), UC (Shen et al., 2018), NSAID-induced enteropathy (Montalto et al., 2013; Syer and Wallace, 2014; Rekatsina et al., 2020) and ulcer healing (Blackler et al., 2015). Microbiome heterogeneity may also influence drug response (Structure, function and diversity of the healthy human microbiome, 2012; Forslund et al., 2015; Wu et al., 2017; Ma et al., 2019; Sharma et al., 2019; Doestzada et al., 2018; Zimmermann et al., 2019; Yip and Chan, 2015; Kashyap et al., 2017).

NSAID disposition, therapeutic efficacy and toxicity are influenced by dynamic and complex host-gut microbiota interactions. Gut microbiota may directly modify NSAID chemistry or manipulate host metabolic processes affecting drug pharmacokinetics and pharmacodynamics. NSAIDs may modify gut microbiota composition resulting in dysbiosis (Maseda and Ricciotti, 2020). Rogers and Aronoff (2016) reported differences in gut microbiota profiles between non-users and NSAID-users. Gut microbiome composition also varied with the type of NSAID ingested (Rogers and Aronoff, 2016).

The capacity to intentionally manipulate the microbiota through diet (Dzutsev et al., 2017; Thiele et al., 2017; Ma et al., 2019), antibiotics/antimicrobials (Lanas and Scarpignato, 2006), fecal microbiota transplant (FMT) (Garber, 2015) or administration of selected probiotic strains (Gionchetti et al., 2000; Ulisse et al., 2001; Gionchetti et al., 2003; Montalto et al., 2010; Montalto et al., 2013; Suzuki et al., 2017; Mortensen et al., 2019; Rekatsina et al., 2020) to enhance efficacy and preserve functional mucosal integrity provides potential therapeutic value for various GI diseases. Recent developments in metaproteomics-based assays offer novel insights into microbiome absolute abundance and functional responses to drugs (Li and et al., 2020). Deployment of next generation sequencing technologies to identify predictive, diagnostic, and prognostic biomarkers need to be further investigated to determine their role in NSAID-induced GI pathology.

FUTURE OPPORTUNITIES AND CLINICAL IMPLEMENTATION

Further research into genetic risk factors predisposing individuals to NSAID-induced upper GI toxicity is justified. Knowledge gaps remain. Currently, there is a shortfall of genome-wide studies, with candidate gene approaches dominating. Lack of consensus and scarcity of independent replication to validate findings has been problematic in published studies. Given the burden and prevalence of NSAID-induced GI ADRs, this is perplexing. A genome-wide approach would allow for unbiased identification of common and rare variants in both PK and PD related genes but would require a large sample size with detailed phenotyping of cases and controls. Indeed, a variability in phenotype definitions is a further challenge in identifying tractable genetic associations. Thus, focused collaborative efforts to standardize phenotype definitions, as shown with other ADRs (Carr et al., 2017; Pirmohamed et al., 2011; Alfirevic et al., 2014; Nicoletti and et al., 2020; Aithal et al., 2011; Behr et al., 2012; Pirmohamed et al., 2011) should be encouraged, drawing inspiration from an adapted 'consensus approach' (Pirmohamed et al., 2011).

CONCLUSION

The pathogenesis of NSAID-induced GI toxicity is complex. PGx is a useful tool to improve pharmacotherapy, and aims to shift the paradigm of dosing regimens being extrapolated to entire populations (Jaccard et al., 2020). Interrogation of genetic factors which predispose individuals to NSAID-induced upper GI toxicity offers unquestionable benefit: pre-emptive testing minimizes the risks of ADRs, informing clinical practice guidelines and therapeutic recommendations to enhance pharmacotherapy (Macías et al., 2020; Theken et al., 2020). To date, the *CYP2C* locus has been shown to be important in some but not all studies. Taken together with the fact that some NSAIDs are only partially metabolized by CYP2C9, but still cause PUD, it can be concluded that *CYP2C* genetic variants are neither necessary nor sufficient to predispose to NSAID-induced PUD, but the presence of low activity *CYP2C9* alleles in a patient given a CYP2C9-metabolised NSAID may increase the risk of PUD. Further work however is also required to identify novel genetic predisposing factors using genome-wide approaches but this will require larger numbers of well phenotyped patients.

Another factor to consider is that as our population continues to age, the prevalence of multimorbidity will increase (Uijen and van de Lisdonk, 2008), which will be accompanied by polypharmacy. Multimorbidity and polypharmacy increase the likelihood of drug- and gene-based interactions (Turner and et al., 2020). The elderly are liable to be prescribed NSAIDs, including LDA, because of the high prevalence of cardiovascular disease and arthritis. The risk of drug-drug interactions with concomitantly administered drugs is likely to be increased in the elderly: these may occur at both pharmacokinetic, (e.g. inhibition of CYP2C9 metabolism) and pharmacodynamic, (e.g. use of NSAIDs and corticosteroids) levels, and may thus contribute to the increased risk of upper GI toxicity. Thus with prolific NSAID use in the over 65 year olds, combinations of risk factors will be common and cumulative, increasing the risks of NSAID-induced ADRs (Laine et al., 2002). It is likely that these ADRs will increase in prevalence despite the use of gastroprotective therapies such as proton pump inhibitors.

In summary therefore, there is a continuing need to define genetic predisposing factors for NSAID-induced PUD because NSAIDs are very widely used, the occurrence of PUD is associated with a high degree of morbidity and mortality, and it is likely with the change in our age demographics, the problems of NSAID-induced PUD are likely to increase rather than decrease.

AUTHOR CONTRIBUTIONS

DC, LM, and MP all contributed to the concept, content, writing and editing of this manuscript.

FUNDING

The work of MP and DC was part-funded by the Medical Research Council grant for the Center for Drug Safety Science, University of Liverpool (Grant Number: MR/L006758/1).

REFERENCES

Agúndez, J. A., Blanca, M., Cornejo-García, J. A., and García-Martín, E. (2015). Pharmacogenomics of Cyclooxygenases. *Pharmacogenomics* 16 (5), 501–522. doi:10.2217/pgs.15.6

Agúndez, J. A., García-Martín, E., and Martínez, C. (2009). Genetically Based Impairment in CYP2C8- and CYP2C9-dependent NSAID Metabolism as a Risk Factor for Gastrointestinal Bleeding: Is a Combination of Pharmacogenomics and Metabolomics Required to Improve Personalized Medicine? *Expert Opin. Drug Metab. Toxicol.* 5 (6), 607–620. doi:10.1517/17425250902970998

Åhsberg, K., Höglund, P., Kim, W.-H., and von Holstein, C. S. (2010). Impact of Aspirin, NSAIDs, Warfarin, Corticosteroids and SSRIs on the Site and Outcome of Non-variceal Upper and Lower Gastrointestinal Bleeding. *Scand. J. Gastroenterol.* 45 (12), 1404–1415. doi:10.3109/00365521.2010.510567

Aithal, G. P., Watkins, P. B., Andrade, R. J., Larrey, D., Molokhia, M., Takikawa, H., et al. (2011). Case Definition and Phenotype Standardization in Drug-Induced Liver Injury. *Clin. Pharmacol. Ther.* 89 (6), 806–815. doi:10.1038/clpt.2011.58

Alfirevic, A., Neely, D., Armitage, J., Chinoy, H., Cooper, R. G., Laaksonen, R., et al. (2014). Phenotype Standardization for Statin-Induced Myotoxicity. *Clin. Pharmacol. Ther.* 96 (4), 470–476. doi:10.1038/clpt.2014.121

Anglin, R., Yuan, Y., Moayyedi, P., Tse, F., Armstrong, D., and Leontiadis, G. I. (2014). Risk of Upper Gastrointestinal Bleeding with Selective Serotonin Reuptake Inhibitors with or without Concurrent Nonsteroidal Anti-inflammatory Use: a Systematic Review and Meta-Analysis. *Am. J. Gastroenterol.* 109 (6), 811–819. doi:10.1038/ajg.2014.82

Arisawa, T, Tahara, T, Shibata, T, Nagasaka, M, Nakamura, M, Kamiya, Y, et al. (2007). Association between Genetic Polymorphisms in the Cyclooxygenase-1 Gene Promoter and Peptic Ulcers in Japan. *Int. J. Mol. Med.* 20 (3), 373–8.

Armstrong, C. P., and Blower, A. L. (1987). Non-steroidal Anti-inflammatory Drugs and Life Threatening Complications of Peptic Ulceration. *Gut* 28 (5), 527–532. doi:10.1136/gut.28.5.527

Bashir, A. I. J., Kankipati, C. S., Jones, S., Newman, R. M., Safrany, S. T., Perry, C. J., et al. (2019). A Novel Mechanism for the Anticancer Activity of Aspirin and Salicylates. *Int. J. Oncol.* 54 (4), 1256–1270. doi:10.3892/ijo.2019.4701

Behr, E. R., January, C., Schulze-Bahr, E., Grace, A. A., Kääb, S., Fiszman, M., et al. (2012). The International Serious Adverse Events Consortium (iSAEC) phenotype standardization project for drug-induced torsades de pointes. *Eur. Heart J.* 34 (26), 1958–1963. doi:10.1093/eurheartj/ehs172

Bhala, N., Bhala, N., Emberson, J., Merhi, A., Abramson, S., Arber, N., et al. (2013). Vascular and Upper Gastrointestinal Effects of Non-steroidal Anti-inflammatory Drugs: Meta-Analyses of Individual Participant Data from Randomised Trials. *Lancet* 382 (9894), 769–779. doi:10.1016/S0140-6736(13)60900-9

Bhandari, P., Bateman, A. C., Mehta, R. L., and Patel, P. (2005). Mucosal Expression of Cyclooxygenase Isoforms 1 and 2 Is Increased with Worsening Damage to the Gastric Mucosa. *Histopathology* 46 (3), 280–286. doi:10.1111/j.1365-2559.2005.02053.x

Bhatt, D. L., Scheiman, J., Abraham, N. S., Antman, E. M., Chan, F. K., Furberg, C. D., et al. (2008). ACCF/ACG/AHA 2008 Expert Consensus Document on Reducing the Gastrointestinal Risks of Antiplatelet Therapy and NSAID Use: a Report of the American College of Cardiology Foundation Task Force on Clinical Expert Consensus Documents. *J. Am. Coll. Cardiol.* 52 (18), 1502–1517. doi:10.1016/j.jacc.2008.08.002

Biondi-Zoccai, G. G. L., Lotrionte, M., Agostoni, P., Abbate, A., Fusaro, M., Burzotta, F., et al. (2006). A Systematic Review and Meta-Analysis on the Hazards of Discontinuing or Not Adhering to Aspirin Among 50 279 Patients at Risk for Coronary Artery Disease. *Eur. Heart J.* 27 (22), 2667–2674. doi:10.1093/eurheartj/ehl334

Bjarnason, I. (2013). Gastrointestinal Safety of NSAIDs and Over-the-counter Analgesics. *Int. J. Clin. Pract.* 67 (s178), 37–42. doi:10.1111/ijcp.12048

Bjarnason, I., Scarpignato, C., Holmgren, E., Olszewski, M., Rainsford, K. D., and Lanas, A. (2018). Mechanisms of Damage to the Gastrointestinal Tract from Nonsteroidal Anti-inflammatory Drugs. *Gastroenterology* 154 (3), 500–514. doi:10.1053/j.gastro.2017.10.049

Bjorkman, D. (1998). Nonsteroidal Anti-inflammatory Drug-Associated Toxicity of the Liver, Lower Gastrointestinal Tract, and Esophagus. *Am. J. Med.* 105 (5Suppl. 1), 17S–21S. doi:10.1016/s0002-9343(98)00276-9

Blackler, R. W., Motta, J.-P., Manko, A., Workentine, M., Bercik, P., Surette, M. G., et al. (2015). Hydrogen Sulphide Protects against NSAID-Enteropathy through Modulation of Bile and the Microbiota. *Br. J. Pharmacol.* 172 (4), 992–1004. doi:10.1111/bph.12961

Blanco, G., Martínez, C., García-Martín, E., and Agúndez, J. A. G. (2005). Cytochrome P450 Gene Polymorphisms and Variability in Response to NSAIDs. *Clin. Res. Regul. Aff.* 22 (2), 57–81. doi:10.1080/10601330500214559

Blanco, G., Martínez, C., Ladero, J. M., Garcia-Martin, E., Taxonera, C., Gamito, F. G., et al. (2008). Interaction of CYP2C8 and CYP2C9 Genotypes Modifies the Risk for Nonsteroidal Anti-inflammatory Drugs-Related Acute Gastrointestinal Bleeding. *Pharmacogenet Genomics* 18 (1), 37–43. doi:10.1097/fpc.0b013e3282f305a9

Blower, A. L., Brooks, A., Fenn, GC., Hill, A., Pearce, MY., Morant, S., et al. (1997). Emergency Admissions for Upper Gastrointestinal Disease and Their Relation to NSAID Use. *Aliment. Pharmacol. Ther.* 11 (2). doi:10.1046/j.1365-2036.1997.d01-604.x

Bombardier, C., Laine, L., Reicin, A., Shapiro, D., Burgos-Vargas, R., Davis, B., et al. (2000). Comparison of Upper Gastrointestinal Toxicity of Rofecoxib and Naproxen in Patients with Rheumatoid Arthritis. *N. Engl. J. Med.* 343 (21), 1520–1528. doi:10.1056/nejm200011233432103

Brune, K., Glatt, M., and Graf, P. (1976). Mechanisms of Action of Anti-inflammatory Drugs. *Gen. Pharmacol. Vasc. Syst.* 7 (1), 27–IN1. doi:10.1016/0306-3623(76)90028-8

Brune, K., and Graf, P. (1978). Non-steroid Anti-inflammatory Drugs: Influence of Extra-cellular pH on Biodistribution and Pharmacological Effects. *Biochem. Pharmacol.* 27 (4), 525–530. doi:10.1016/0006-2952(78)90388-x

Butt, J. H., Barthel, J. S., and Moore, R. A. (1988). Clinical Spectrum of the Upper Gastrointestinal Effects of Nonsteroidal Anti-inflammatory Drugs. *Am. J. Med.* 84 (2a), 5–14. doi:10.1016/0002-9343(88)90248-3

Calatayud, S., and Esplugues, J. V. (2016). "Chemistry, Pharmacodynamics, and Pharmacokinetics of NSAIDs," in *NSAIDs and Aspirin: Recent Advances and Implications for Clinical Management*. Editor A. Lanas (Cham: Springer International Publishing), 3–16. doi:10.1007/978-3-319-33889-7_1

Campbell, K., and Steele, R. J. (1991). Non-steroidal Anti-inflammatory Drugs and Complicated Diverticular Disease: a Case-Control Study. *Br. J. Surg.* 78 (2), 190–191. doi:10.1002/bjs.1800780218

Carbonell, N., Verstuyft, C., Massard, J., Letierce, A., Cellier, C., Deforges, L., et al. (2010). CYP2C9*3 Loss-Of-Function Allele Is Associated with Acute Upper Gastrointestinal Bleeding Related to the Use of NSAIDs Other Than Aspirin. *Clin. Pharmacol. Ther.* 87 (6), 693–698. doi:10.1038/clpt.2010.33

Carr, D. F., Ayehunie, S., Davies, A., Duckworth, C. A., French, S., Hall, N., et al. (2017). Towards Better Models and Mechanistic Biomarkers for Drug-Induced Gastrointestinal Injury. *Pharmacol. Ther.* 172, 181–194. doi:10.1016/j.pharmthera.2017.01.002

Castellsague, J., Riera-Guardia, N., Calingaert, B., Varas-Lorenzo, C., Fourrier-Reglat, A., Nicotra, F., et al. (2012). Individual NSAIDs and Upper Gastrointestinal Complications. *Drug Saf.* 35 (12), 1127–1146. doi:10.1007/bf03261999

Catella-Lawson, F., Reilly, M. P., Kapoor, S. C., Cucchiara, A. J., DeMarco, S., Tournier, B., et al. (2001). Cyclooxygenase Inhibitors and the Antiplatelet Effects of Aspirin. *N. Engl. J. Med.* 345 (25), 1809–1817. doi:10.1056/nejmoa003199

Cayla, G., Collet, J.-P., Silvain, J., Thiefin, G., Woimant, F., and Montalescot, G. (2012). Prevalence and Clinical Impact of Upper Gastrointestinal Symptoms in Subjects Treated with Low Dose Aspirin: The UGLA Survey. *Int. J. Cardiol.* 156 (1), 69–75. doi:10.1016/j.ijcard.2010.10.027

Chan, F. K., Lanas, A., Scheiman, J., Berger, M. F., Nguyen, H., and Goldstein, J. L. (2010). Celecoxib versus Omeprazole and Diclofenac in Patients with Osteoarthritis and Rheumatoid Arthritis (CONDOR): a Randomised Trial. *The Lancet* 376 (9736), 173–179. doi:10.1016/s0140-6736(10)60673-3

Chan, F. K. L., Ching, J. Y. L., Tse, Y. K., Lam, K., Wong, G. L. H., Ng, S. C., et al. (2017). Gastrointestinal Safety of Celecoxib versus Naproxen in Patients with Cardiothrombotic Diseases and Arthritis after Upper Gastrointestinal Bleeding (CONCERN): an Industry-independent, Double-Blind, Double-Dummy, Randomised Trial. *The Lancet* 389 (10087), 2375–2382. doi:10.1016/s0140-6736(17)30981-9

Chan, F. K. L., Hung, L. C. T., Suen, B. Y., Wong, V. W. S., Hui, A. J., Wu, J. C. Y., et al. (2005). *Effect of Celecoxib on the Healing of Complicated Gastric Ulcers: A Prospective, Double Blinded, Randomized Trial*, A24–A25.

Chi, T.-Y., Zhu, H.-M., and Zhang, M. (2018). Risk Factors Associated with Nonsteroidal Anti-inflammatory Drugs (NSAIDs)-Induced Gastrointestinal Bleeding Resulting on People over 60 Years Old in Beijing. *Medicine* 97 (18), e0665. doi:10.1097/md.0000000000010665

Cho, J. H., Choi, J. S., Chun, S. W., Lee, S., Han, K. J., Kim, H, M., et al. (2016). The IL-1B Genetic Polymorphism Is Associated with Aspirin-Induced PepticUlcers in a Korean Ethnic Group. *Gut Liver* 10 (3), 362–368. doi:10.5009/gnl15129

Cohen, M.M. (1987). Role of Endogenous Prostaglandins in Gastric Secretion and Mucosal Defense. *Clin. Invest. Med.* 10 (3), 226–31.

Collaboration, A. T.Antithrombotic Trialists' Collaboration (2002). Collaborative Meta-Analysis of Randomised Trials of Antiplatelet Therapy for Prevention of Death, Myocardial Infarction, and Stroke in High Risk Patients. *Bmj* 324 (7329), 71–86. doi:10.1136/bmj.324.7329.71

Cryer, B. (2004). COX-2-specific Inhibitor or Proton Pump Inhibitor Plus Traditional NSAID: Is Either Approach Sufficient for Patients at Highest Risk of NSAID-Induced Ulcers? *Gastroenterology* 127 (4), 1256–1258. doi:10.1053/j.gastro.2004.08.029

Dai, D., Zeldin, D. C., Blaisdell, J. A., Chanas, B., Coulter, S. J., Ghanayem, B. I., et al. (2001). Polymorphisms in Human CYP2C8 Decrease Metabolism of the Anticancer Drug Paclitaxel and Arachidonic Acid. *Pharmacogenetics* 11 (7), 597–607. doi:10.1097/00008571-200110000-00006

Dajani, E. Z., and Agrawal, N. M. (1995). Prevention of Nonsteroidal Anti-inflammatory Drug-Induced Gastroduodenal Ulcers: Role of Mucosal Protective and Gastric Antisecretory Drugs. *Dig. Dis.* 13 (Suppl. 1), 48–61. doi:10.1159/000171526

Davies, N. M., McLachlan, A. J., Day, R. O., and Williams, K. M. (2000). Clinical Pharmacokinetics and Pharmacodynamics of Celecoxib. *Clin. Pharmacokinet.* 38 (3), 225–242. doi:10.2165/00003088-200038030-00003

Davies, N. M., and Wallace, J. L. (1996). Selective Inhibitors of Cyclooxygenase-2. *Drugs and Aging* 9 (6), 406–417. doi:10.2165/00002512-199609060-00004

de Abajo, F. J., Rodríguez, L. A. G., and Montero, D. (1999). Association between Selective Serotonin Reuptake Inhibitors and Upper Gastrointestinal Bleeding: Population Based Case-Control Study. *Bmj* 319 (7217), 1106–1109. doi:10.1136/bmj.319.7217.1106

de Jong, J. C. F., Van Den Berg, P. B., Tobi, H., and De Jong Van Den Berg, L. T. W. (2003). Combined Use of SSRIs and NSAIDs Increases the Risk of

Gastrointestinal Adverse Effects. *Br. J. Clin. Pharmacol.* 55 (6), 591–595. doi:10.1046/j.0306-5251.2002.01770.x

Dickman, A., and Ellershaw, J. (2004). For Discussion NSAIDs: Gastroprotection or Selective COX-2 Inhibitor? *Palliat. Med.* 18 (4), 275–286. doi:10.1191/0269216304pm894fd

Doestzada, M., Vila, A. V., Zhernakova, A., Koonen, D. P. Y., Weersma, R. K., Touw, D. J., et al. (2018). Pharmacomicrobiomics: a Novel Route towards Personalized Medicine? *Protein Cell* 9 (5), 432–445. doi:10.1007/s13238-018-0547-2

Dunlap, J. J., and Patterson, S. (2019). Peptic Ulcer Disease. *Gastroenterol. Nurs.* 42 (5), 451–454. doi:10.1097/sga.0000000000000478

Dzutsev, A., Badger, J. H., Perez-Chanona, E., Roy, S., Salcedo, R., Smith, C. K., et al. (2017). Microbes and Cancer. *Annu. Rev. Immunol.* 35, 199–228. doi:10.1146/annurev-immunol-051116-052133

Eichelbaum, M., Ingelman-Sundberg, M., and Evans, W. E. (2006). Pharmacogenomics and Individualized Drug Therapy. *Annu. Rev. Med.* 57, 119–137. doi:10.1146/annurev.med.56.082103.104724

Evans, W. E., and Relling, M. V. (1999). Pharmacogenomics: Translating Functional Genomics into Rational Therapeutics. *Science* 286 (5439), 487–491. doi:10.1126/science.286.5439.487

Faucheron, J.-L. (1999). Toxicity of Non-steroidal Anti-inflammatory Drugs in the Large Bowel. *Eur. J. Gastroenterol. Hepatol.* 11 (4), 389–392. doi:10.1097/00042737-199904000-00005

Figueiras, A., Estany-Gestal, A., Aguirre, C., Ruiz, B., Vidal, X., Carvajal, A., et al. (2016). CYP2C9 Variants as a Risk Modifier of NSAID-Related Gastrointestinal Bleeding: a Case–Control Study. *Pharmacogenetics and Genomics* 26 (2). doi:10.1097/fpc.0000000000000186

Forrest, E. H., Russell, R. I., Orchard, T. R., and Jewell, D. P. (1999). Peripheral Arthropathies in Inflammatory Bowel Disease Reply. *Gut* 44 (3), 439. doi:10.1136/gut.44.3.439

Forslund, K., au, fnm., Hildebrand, F., Nielsen, T., Falony, G., Le Chatelier, E., et al. (2015). Disentangling Type 2 Diabetes and Metformin Treatment Signatures in the Human Gut Microbiota. *Nature* 528 (7581), 262–266. doi:10.1038/nature15766

Fries, J. F., Williams, C. A., Bloch, D. A., and Michel, B. A. (1991). Nonsteroidal Anti-inflammatory Drug-Associated Gastropathy: Incidence and Risk Factor Models. *Am. J. Med.* 91 (3), 213–222. doi:10.1016/0002-9343(91)90118-h

Fries, J. (1996). Toward an Understanding of NSAID-Related Adverse Events: the Contribution of Longitudinal Data. *Scand. J. Rheumatol.* 25, 3–8. doi:10.3109/03009749609097225

Funk, C. D., Funk, L. B., Kennedy, M. E., Pong, A. S., and Fitzgerald, G. A. (1991). Human Platelet/erythroleukemia Cell Prostaglandin G/H Synthase: cDNA Cloning, Expression, and Gene Chromosomal Assignment. *FASEB j.* 5 (9), 2304–2312. doi:10.1096/fasebj.5.9.1907252

Gabriel, S. E., Jaakkimainen, L., and Bombardier, C. (1991). Risk for Serious Gastrointestinal Complications Related to Use of Nonsteroidal Anti-inflammatory Drugs. *Ann. Intern. Med.* 115 (10), 787–796. doi:10.7326/0003-4819-115-10-787

Gaedigk, A., Ingelman-Sundberg, M., Miller, N. A., Leeder, J. S., Whirl-Carrillo, M., and Klein, T. E. (2018). The Pharmacogene Variation (PharmVar) Consortium: Incorporation of the Human Cytochrome P450 (CYP) Allele Nomenclature Database. *Clin. Pharmacol. Ther.* 103 (3), 399–401. doi:10.1002/cpt.910

Garber, K. (2015). Drugging the Gut Microbiome. *Nat. Biotechnol.* 33 (3), 228–231. doi:10.1038/nbt.3161

Garcia Rodriguez, L., and Hernández-Díaz, S. (2001). The Risk of Upper Gastrointestinal Complications Associated with Nonsteroidal Anti-inflammatory Drugs, Glucocorticoids, Acetaminophen, and Combinations of These Agents. *Arthritis Res.* 3, 98–101. doi:10.1186/ar146

García Rodríguez, L., and Jick, H. (1994). Risk of Upper Gastrointestinal Bleeding and Perforation Associated with Individual Non-steroidal Anti-inflammatory Drugs. *The Lancet* 343 (8900), 769–772. doi:10.1016/s0140-6736(94)91843-0

García-Martín, E., Martínez, C., Ladero, J. M., and Agúndez, J. A. G. (2006). Interethnic and Intraethnic Variability of CYP2C8 and CYP2C9 Polymorphisms in Healthy Individuals. *Mol. Diag. Ther.* 10 (1), 29–40. doi:10.1007/bf03256440

García-Rayado, G., Navarro, M., and Lanas, A. (2018). NSAID Induced Gastrointestinal Damage and Designing GI-Sparing NSAIDs. *Expert Rev. Clin. Pharmacol.* 11 (10), 1031–1043. doi:10.1080/17512433.2018.1516143

García-Martín, E., Martínez, C., Tabarés, B., Frías, J., and Agúndez, J. A. G. (2004). Interindividual Variability in Ibuprofen Pharmacokinetics Is Related to Interaction of Cytochrome P450 2C8 and 2C9 Amino Acid Polymorphisms*1. *Clin. Pharmacol. Ther.* 76 (2), 119–127. doi:10.1016/j.clpt.2004.04.006

Gargallo, C. J., Sostres, C., and Lanas, A. (2014). Prevention and Treatment of NSAID Gastropathy. *Curr. Treat. Options. Gastro*. 12 (4), 398–413. doi:10.1007/s11938-014-0029-4

Geis, GS, Stead, H, Wallemark, CB, and Nicholson, PA (1991). Prevalence of Mucosal Lesions in the Stomach and Duodenum Due to Chronic Use of NSAID in Patients with Rheumatoid Arthritis or Osteoarthritis, and Interim Report on Prevention by Misoprostol of Diclofenac Associated Lesions. *J. Rheumatol. Suppl.* 28, 11–4.

Gevers, D., Kugathasan, S., Denson, L. A., Vázquez-Baeza, Y., Van Treuren, W., Ren, B., et al. (2014). The Treatment-Naive Microbiome in New-Onset Crohn's Disease. *Cell Host and Microbe* 15 (3), 382–392. doi:10.1016/j.chom.2014.02.005

Gionchetti, P., Rizzello, F., Helwig, U., Venturi, A., Lammers, K. M., Brigidi, P., et al. (2003). Prophylaxis of Pouchitis Onset with Probiotic Therapy: a Double-Blind, Placebo-Controlled Trial. *Gastroenterology* 124 (5), 1202–1209. doi:10.1016/s0016-5085(03)00171-9

Gionchetti, P., Rizzello, F., Venturi, A., Brigidi, P., Matteuzzi, D., Bazzocchi, G., et al. (2000). Oral Bacteriotherapy as Maintenance Treatment in Patients with Chronic Pouchitis: a Double-Blind, Placebo-Controlled Trial. *Gastroenterology* 119 (2), 305–309. doi:10.1053/gast.2000.9370

Goldstein, J. A., and de Morais, S. M. F. (1994). Biochemistry and Molecular Biology of the Human CYP2C Subfamily. *Pharmacogenetics* 4 (6), 285–300. doi:10.1097/00008571-199412000-00001

Goldstein, J., and Cryer, B. (2015). Gastrointestinal Injury Associated with NSAID Use: a Case Study and Review of Risk Factors and Preventative Strategies. *Dhps* 7, 31–41. doi:10.2147/dhps.s71976

Goldstein, J. L., Chan, F. K. L., Lanas, A., Wilcox, C. M., Peura, D., Sands, G. H., et al. (2011). Haemoglobin Decreases in NSAID Users over Time: an Analysis of Two Large Outcome Trials. *Aliment. Pharmacol. Ther.* 34 (7), 808–816. doi:10.1111/j.1365-2036.2011.04790.x

Goldstein, J. L., and Lefkowith, J. B. (1998). Public Misunderstanding of Nonsteroidal Antiinflammatory Drug (NSAID)-mediated Gastrointestinal (GI) Toxicity: a Serious Potential Health Threat. *Gastroenterology* 114, A136. doi:10.1016/s0016-5085(98)80552-0

Gómez-Acebo, I., Dierssen-Sotos, T., de Pedro, M., Pérez-Gómez, B., Castaño-Vinyals, G., Fernández-Villa, F., et al. (2018). Epidemiology of Non-steroidal Anti-inflammatory Drugs Consumption in Spain. The MCC-Spain Study. *BMC Public Health* 18 (1), 1134. doi:10.1186/s12889-018-6019-z

González-Pérez, A., Sáez, M. E., Johansson, S., Nagy, P., and García Rodríguez, L. A (2014). Risk Factors Associated with Uncomplicated Peptic Ulcer and Changes in Medication Use after Diagnosis. *PloS one* 9 (7), e101768. doi:10.1371/journal.pone.0101768

Gretzer, B., Maricic, N., Respondek, M., Schuligoi, R., and Peskar, B. M. (2001). Effects of Specific Inhibition of Cyclo-Oxygenase-1 and Cyclo-Oxygenase-2 in the Rat Stomach with normal Mucosa and after Acid challenge. *Br. J. Pharmacol.* 132 (7), 1565–1573. doi:10.1038/sj.bjp.0703955

Griffin, M. R., Piper, J. M., Daugherty, J. R., Snowden, M., and Ray, W. A. (1991). Nonsteroidal Anti-inflammatory Drug Use and Increased Risk for Peptic Ulcer Disease in Elderly Persons. *Ann. Intern. Med.* 114 (4), 257–263. doi:10.7326/0003-4819-114-4-257

Guengerich, F. P. (2020). A history of the roles of cytochrome P450 enzymes in the toxicity of drugs. *Toxicol. Res.* 37, 1–23. doi:10.1007/s43188-020-00056-z

Gutthann, S. P., GarcíaRodríguez, L. A., and Raiford, D. S. (1997). Individual Nonsteroidal Antiinflammatory Drugs and Other Risk Factors for Upper Gastrointestinal Bleeding and Perforation. *Epidemiology* 8 (1), 18–24. doi:10.1097/00001648-199701000-00003

Halushka, M., Walker, L. P., and Halushka, P. V. (2003). Genetic Variation in Cyclooxygenase 1: Effects on Response to Aspirin. *Clin. Pharmacol. Ther.* 73 (1), 122–130. doi:10.1067/mcp.2003.1

Handa, O., Majima, A., Onozawa, Y., Horie, H., Uehara, Y., Fukui, A., et al. (2014). The Role of Mitochondria-Derived Reactive Oxygen Species in the Pathogenesis of Non-steroidal Anti-inflammatory Drug-Induced Small Intestinal Injury. *Free Radic. Res.* 48 (9), 1095–1099. doi:10.3109/10715762.2014.928411

Hawkey, C. J., and Langman, M. J. (2003). Non-steroidal Anti-inflammatory Drugs: Overall Risks and Management. Complementary Roles for COX-2 Inhibitors and Proton Pump Inhibitors. *Gut* 52 (4), 600–608. doi:10.1136/gut.52.4.600

Helin-Salmivaara, A., Huttunen, T., Grönroos, J. M., Klaukka, T., and Huupponen, R. (2007). Risk of Serious Upper Gastrointestinal Events with Concurrent Use of NSAIDs and SSRIs: a Case-Control Study in the General Population. *Eur. J. Clin. Pharmacol.* 63 (4), 403–408. doi:10.1007/s00228-007-0263-y

Henry, D., Dobson, A., and Turner, C. (1993). Variability in the Risk of Major Gastrointestinal Complications from Nonaspirin Nonsteroidal Anti-inflammatory Drugs. *Gastroenterology* 105 (4), 1078–1088. doi:10.1016/0016-5085(93)90952-9

Herlitz, J., Tóth, P. P., and Næsdal, J. (2010). Low-Dose Aspirin Therapy for Cardiovascular Prevention. *Am. J. Cardiovasc. Drugs* 10 (2), 125–141. doi:10.2165/11318440-000000000-00000

Hernández-Díaz, S., and Rodríguez, L. A. G. (2000). Association between Nonsteroidal Anti-inflammatory Drugs and Upper Gastrointestinal Tract Bleeding/Perforation. *Arch. Intern. Med.* 160 (14), 2093–2099. doi:10.1001/archinte.160.14.2093

Hirschowitz, BI (1994). Nonsteroidal Antiinflammatory Drugs and the Gastrointestinal Tract. *Gastroenterologist* 2 (3), 207–23.

Huang, J.-Q., Sridhar, S., and Hunt, R. H. (2002). Role of *Helicobacter pylori* Infection and Non-steroidal Anti-inflammatory Drugs in Peptic-Ulcer Disease: a Meta-Analysis. *The Lancet* 359 (9300), 14–22. doi:10.1016/s0140-6736(02)07273-2

Human Microbiome Project Consortium (2012). Structure, Function and Diversity of the Healthy Human Microbiome. *Nature* 486 (7402), 207–214. doi:10.1038/nature11234

Ingelman-Sundberg, M., Sim, S. C., Gomez, A., and Rodriguez-Antona, C. (2007). Influence of Cytochrome P450 Polymorphisms on Drug Therapies: Pharmacogenetic, Pharmacoepigenetic and Clinical Aspects. *Pharmacol. Ther.* 116 (3), 496–526. doi:10.1016/j.pharmthera.2007.09.004

Ishihara, M., Ohmiya, N., Nakamura, M., Funasaka, K., Miyahara, R., Ohno, E., et al. (2014). Risk Factors of Symptomatic NSAID-Induced Small Intestinal Injury and Diaphragm Disease. *Aliment. Pharmacol. Ther.* 40 (5), 538–547. doi:10.1111/apt.12858

Jaccard, E, Redin, C, Girardin, F, Waeber, G, Fellay, J, and Vollenweider, P (2020). Pharmacogenomics : a Toolbox to Improve Drug Prescription. *Rev. Med. Suisse* 16 (716), 2259–2263.

Jankovic, SM, Aleksic, J, Rakovic, S, Aleksic, A, Stevanovic, I, Stefanovic-Stoimenov, N, et al. (2009). Nonsteroidal Antiinflammatory Drugs and Risk of Gastrointestinal Bleeding Among Patients on Hemodialysis. *J. Nephrol.* 22 (4), 502–7.

Johansson, I., and Ingelman-Sundberg, M. (2011). Genetic Polymorphism and Toxicology-With Emphasis on Cytochrome P450. *Toxicol. Sci.* 120 (1), 1–13. doi:10.1093/toxsci/kfq374

Johnsen, SP, Sørensen, HT, Mellemkjoer, L, Blot, WJ, Nielsen, GL, McLaughlin, JK, et al. (2001). Hospitalisation for Upper Gastrointestinal Bleeding Associated with Use of Oral Anticoagulants. *Thromb. Haemost.* 86 (2), 563–8.

Kapur, B. M., Lala, P. K., and Shaw, J. L. (2014). Pharmacogenetics of Chronic Pain Management. *Clin. Biochem.* 47 (13-14), 1169–1187. doi:10.1016/j.clinbiochem.2014.05.065

Kashyap, P. C., Chia, N., Nelson, H., Segal, E., and Elinav, E. (2017). Microbiome at the Frontier of Personalized Medicine. *Mayo Clinic Proc.* 92 (12), 1855–1864. doi:10.1016/j.mayocp.2017.10.004

Kim, J., Lee, J., Shin, C. M., Lee, D. H., and Park, B.-J. (2015). Risk of Gastrointestinal Bleeding and Cardiovascular Events Due to NSAIDs in the Diabetic Elderly Population. *BMJ Open Diab. Res. Care*, 3(1): p. e000133. doi:10.1136/bmjdrc-2015-000133

Kirchheiner, J., Meineke, I., Freytag, G., Meisel, C., Roots, I., and Brockmöller, J. (2002). Enantiospecific Effects of Cytochrome P450 2C9 Amino Acid Variants on Ibuprofen Pharmacokinetics and on the Inhibition of Cyclooxygenases 1 and 2*. *Clin. Pharmacol. Ther.* 72 (1), 62–75. doi:10.1067/mcp.2002.125726

Krasniqi, V., Dimovski, A., Domjanović, I. K., Bilić, I., and Božina, N. (2016). How Polymorphisms of the Cytochrome P450 Genes Affect Ibuprofen and Diclofenac Metabolism and Toxicity/Kako Polimorfizmi Gena Citokroma P450 Utječu Na Metabolizam I Toksičnost Ibuprofena I Diklofenaka. *Arhiv za higijenu rada i toksikologiju* 67 (1), 1–8. doi:10.1515/aiht-2016-67-2754

Kuehl, G. E., Lampe, J. W., Potter, J. D., and Bigler, J. (2005). Glucuronidation of Nonsteroidal Anti-inflammatory Drugs: Identifying the Enzymes Responsible in Human Liver Microsomes. *Drug Metab. Dispos.* 33 (7), 1027–1035. doi:10.1124/dmd.104.002527

Laine, L. (2001). Approaches to Nonsteroidal Anti-inflammatory Drug Use in the High-Risk Patient. *Gastroenterology* 120 (3), 594–606. doi:10.1053/gast.2001.21907

Laine, L., Bombardier, C., Hawkey, C. J., Davis, B., Shapiro, D., Brett, C., et al. (2002). Stratifying the Risk of NSAID-Related Upper Gastrointestinal Clinical Events: Results of a Double-Blind Outcomes Study in Patients with Rheumatoid Arthritis. *Gastroenterology* 123 (4), 1006–1012. doi:10.1053/gast.2002.36013

Laine, L. (2006). GI Risk and Risk Factors of NSAIDs. *J. Cardiovasc. Pharmacol.* 47, S60–S66. doi:10.1097/00005344-200605001-00011

Lanas, A., García-Rodríguez, L. A., Arroyo, M. T., Bujanda, L., Gomollón, F., Forné, M., et al. (2007). Effect of Antisecretory Drugs and Nitrates on the Risk of Ulcer Bleeding Associated with Nonsteroidal Anti-inflammatory Drugs, Antiplatelet Agents, and Anticoagulants. *Am. J. Gastroenterol.* 102, 507–515. doi:10.1111/j.1572-0241.2006.01062.x

Lanas, A., and Hunt, R. (2006). Prevention of Anti-inflammatory Drug-induced Gastrointestinal Damage: Benefits and Risks of Therapeutic Strategies. *Ann. Med.* 38 (6), 415–428. doi:10.1080/07853890600925843

Lanas, A., and Scarpignato, C. (2006). Microbial flora in NSAID-Induced Intestinal Damage: a Role for Antibiotics? *Digestion* 73 (Suppl. 1), 136–150. doi:10.1159/000089789

Lanas, Á., Carrera-Lasfuentes, P., Arguedas, Y., García, S., Bujanda, L., Calvet, X., et al. (2015). Risk of Upper and Lower Gastrointestinal Bleeding in Patients Taking Nonsteroidal Anti-inflammatory Drugs, Antiplatelet Agents, or Anticoagulants. *Clin. Gastroenterol. Hepatol.* 13 (5), 906–912. doi:10.1016/j.cgh.2014.11.007

Langenbach, R., Morham, S. G., Tiano, H. F., Loftin, C. D., Ghanayem, B. I., Chulada, P. C., et al. (1995). Prostaglandin Synthase 1 Gene Disruption in Mice Reduces Arachidonic Acid-Induced Inflammation and Indomethacin-Induced Gastric Ulceration. *Cell* 83 (3), 483–492. doi:10.1016/0092-8674(95)90126-4

Langman, M. J., Morgan, L., and Worrall, A. (1985). Use of Anti-inflammatory Drugs by Patients Admitted with Small or Large Bowel Perforations and Haemorrhage. *Bmj* 290 (6465), 347–349. doi:10.1136/bmj.290.6465.347

Langman, M. J. S. (1988). Ulcer Complications and Nonsteroidal Anti-inflammatory Drugs. *Am. J. Med.* 84 (2a), 15–19. doi:10.1016/0002-9343(88)90249-5

Langman, M. J. S. (2001). Ulcer Complications Associated with Anti-inflammatory Drug Use. What Is the Extent of the Disease burden? *Pharmacoepidem. Drug Safe.* 10 (1), 13–19. doi:10.1002/pds.561

Lanza, F. L., Chan, F. K. L., and Quigley, E. M. M. (2009). Guidelines for Prevention of NSAID-Related Ulcer Complications. *Am. J. Gastroenterol.* 104 (3), 728–738. doi:10.14309/00000434-200903000-00035

Laporte, J.-R., Ib????ez, L., Vidal, X., Vendrell, L., and Leone, R. (2004). Upper Gastrointestinal Bleeding Associated with the Use of NSAIDs. *Drug Saf.* 27 (6), 411–420. doi:10.2165/00002018-200427060-00005

Larkai, EN, Smith, JL, Lidsky, MD, and Graham, DY (1987). Gastroduodenal Mucosa and Dyspeptic Symptoms in Arthritic Patients during Chronic Nonsteroidal Anti-inflammatory Drug Use. *Am. J. Gastroenterol.* 82 (11), 1153–8.

Larkai, E. N., Smith, J. L., Lidsky, M. D., Sessoms, S. L., and Graham, D. Y. (1989). Dyspepsia in NSAID Users. *J. Clin. Gastroenterol.* 11 (2), 158–162. doi:10.1097/00004836-198904000-00009

Larson, A. M., Polson, J., Fontana, R. J., Davern, T. J., Lalani, E., Hynan, L. S., et al. (2005). Acetaminophen-induced Acute Liver Failure: Results of a United States Multicenter, Prospective Study. *Hepatology* 42 (6), 1364–1372. doi:10.1002/hep.20948

Lazzaroni, M., and Bianchi Porro, G. (2001). Prophylaxis and Treatment of Non-steroidal Anti-inflammatory Drug-Induced Upper Gastrointestinal Side-Effects. *Dig. Liver Dis.* 33 (Suppl. 2), S44–S58. doi:10.1016/s1590-8658(01)80158-4

Lee, C. R., Goldstein, J. A., and Pieper, J. A. (2002). Cytochrome P450 2C9 Polymorphisms: a Comprehensive Review of the *In-Vitro* and Human Data. *Pharmacogenetics* 12 (3), 251–263. doi:10.1097/00008571-200204000-00010

Lee, S. J., Park, M. K., Shin, D., and Chun, M. H. (2017). Variability of the Drug Response to Nonsteroidal Anti-inflammatory Drugs According to

Cyclooxygenase-2 Genetic Polymorphism. *Dddt* 11, 2727–2736. doi:10.2147/dddt.s143807

Leontiadis, G. I., Sreedharan, A., Dorward, S., Barton, P., Delaney, B., Howden, C W., et al. (2007). Systematic Reviews of the Clinical Effectiveness and Cost-Effectiveness of Proton Pump Inhibitors in Acute Upper Gastrointestinal Bleeding. *Health Technol. Assess.* 11 (51), 1–164. doi:10.3310/hta11510

Li, L., Ning, Z., Zhang, X., Mayne, J., Cheng, K., Stintzi, A., et al. (2020). RapidAIM: a Culture- and Metaproteomics-Based Rapid Assay of Individual Microbiome Responses to Drugs. *Microbiome* 8 (1), 33. doi:10.1186/s40168-020-00806-z

Lichtenberger, L. M. (2001). Where Is the Evidence that Cyclooxygenase Inhibition Is the Primary Cause of Nonsteroidal Anti-inflammatory Drug (NSAID)-induced Gastrointestinal Injury? *Biochem. Pharmacol.* 61 (6), 631–637. doi:10.1016/s0006-2952(00)00576-1

Ligumsky, M., Golanska, E. M., Hansen, D. G., and Kauffman, G. L. (1983). Aspirin Can Inhibit Gastric Mucosal Cyclo-Oxygenase without Causing Lesions in Rat. *Gastroenterology* 84 (4), 756–761. doi:10.1016/0016-5085(83)90143-9

Ligumsky, M., Sestieri, M., Karmeli, F., Zimmerman, J., Okon, E., and Rachmilewitz, D. (1990). Rectal Administration of Nonsteroidal Andinflammatory Drugs. *Gastroenterology* 98 (5), 1245–1249. doi:10.1016/0016-5085(90)90340-7, Part 1

Loetsch, J., and Oertel, B. (2013). "NSAIDs, Pharmacokinetics," in *Encyclopedia of Pain*. Editors G. F. Gebhart and R. F. Schmidt (Berlin, Heidelberg: Springer Berlin Heidelberg), 2350–2361. doi:10.1007/978-3-642-28753-4_2853

Lonjou, C., Borot, N., Sekula, P., Ledger, N., Thomas, L., Halevy, S., et al. (2008). A European Study of HLA-B in Stevens-Johnson Syndrome and Toxic Epidermal Necrolysis Related to Five High-Risk Drugs. *Pharmacogenet Genomics* 18 (2), 99–107. doi:10.1097/fpc.0b013e3282f3ef9c

Lopezrodriguez, R., Novalbos, J., Gallegosandin, S., Romanmartinez, M., Torrado, J., Gisbert, J., et al. (2008). Influence of CYP2C8 and CYP2C9 Polymorphisms on Pharmacokinetic and Pharmacodynamic Parameters of Racemic and Enantiomeric Forms of Ibuprofen in Healthy Volunteers. *Pharmacol. Res.* 58 (1), 77–84. doi:10.1016/j.phrs.2008.07.004

Ma, J., Yang, X. Y., Qiao, L., Liang, L. Q., and Chen, M. H. (2008). CYP2C9 Polymorphism in Non-steroidal Anti-inflammatory Drugs-Induced Gastropathy. *J. Dig. Dis.* 9 (2), 79–83. doi:10.1111/j.1751-2980.2008.00326.x

Ma, W., Mao, Q., Xia, W., Dong, G., Yu, C., and Jiang, F. (2019). Gut Microbiota Shapes the Efficiency of Cancer Therapy. *Front. Microbiol.* 10 (1050). doi:10.3389/fmicb.2019.01050

Mace, O. J., Tehan, B., and Marshall, F. (2015). Pharmacology and Physiology of Gastrointestinal Enteroendocrine Cells. *Pharmacol. Res. Perspect.* 3 (4), e00155. doi:10.1002/prp2.155

Macías, Y., Gómez Tabales, J., García-Martín, E., and Agúndez, J. A. G. (2020). An Update on the Pharmacogenomics of NSAID Metabolism and the Risk of Gastrointestinal Bleeding. *Expert Opin. Drug Metab. Toxicol.* 16 (4), 319–332. doi:10.1080/17425255.2020.1744563

Maiden, L., Thjodleifsson, B., Theodors, A., Gonzalez, J., and Bjarnason, I. (2005). A Quantitative Analysis of NSAID-Induced Small Bowel Pathology by Capsule Enteroscopy. *Gastroenterology* 128 (5), 1172–1178. doi:10.1053/j.gastro.2005.03.020

Mallah, N., Zapata-Cachafeiro, M., Aguirre, C., Ibarra-García, E., Palacios-Zabalza, I., Macías-García, F., et al. (2020). Influence of Polymorphisms Involved in Platelet Activation and Inflammatory Response on Aspirin-Related Upper Gastrointestinal Bleeding: A Case-Control Study. *Front. Pharmacol.* 11, 860. doi:10.3389/fphar.2020.01072

Martin, J. H., Begg, E. J., Kennedy, M. A., Roberts, R., and Barclay, M. L. (2001). Is Cytochrome P450 2C9 Genotype Associated with NSAID Gastric Ulceration? *Br. J. Clin. Pharmacol.* 51 (6), 627–630. doi:10.1046/j.0306-5251.2001.01398.x

Martínez, C., Blanco, G., Ladero, J. M., García-Martín, E., Taxonera, C., Gamito, F. G., et al. (2004). Genetic Predisposition to Acute Gastrointestinal Bleeding after NSAIDs Use. *Br. J. Pharmacol.* 141 (2), 205–208. doi:10.1038/sj.bjp.0705623

Masclee, G. M. C., Valkhoff, V. E., van Soest, E. M., Schade, R., Mazzaglia, G., Molokhia, M., et al. (2013). Cyclo-oxygenase-2 Inhibitors or Nonselective NSAIDs Plus Gastroprotective Agents: what to Prescribe in Daily Clinical Practice? *Aliment. Pharmacol. Ther.* 38 (2), 178–189. doi:10.1111/apt.12348

Maseda, D., and Ricciotti, E. (2020). NSAID-gut Microbiota Interactions. *Front. Pharmacol.* 11, 1153. doi:10.3389/fphar.2020.01153

McCarthy, D. M. (1989). Nonsteroidal Antiinflammatory Drug-Induced Ulcers: Management by Traditional Therapies. *Gastroenterology* 96 (2 Pt 2 Suppl. l), 662–674. doi:10.1016/s0016-5085(89)80063-0

McCarthy, D. (1998). Nonsteroidal Anti-inflammatory Drug-Related Gastrointestinal Toxicity: Definitions and Epidemiology. *Am. J. Med.* 105 (5a), 3s–9s. doi:10.1016/s0002-9343(98)00274-5

McGettigan, P., and Henry, D. (2013). Use of Non-steroidal Anti-inflammatory Drugs that Elevate Cardiovascular Risk: an Examination of Sales and Essential Medicines Lists in Low-, Middle-, and High-Income Countries. *Plos Med.* 10 (2), e1001388. doi:10.1371/journal.pmed.1001388

Melcarne, L., García-Iglesias, P., and Calvet, X. (2016). Management of NSAID-Associated Peptic Ulcer Disease. *Expert Rev. Gastroenterol. Hepatol.* 10 (6), 723–733. doi:10.1586/17474124.2016.1142872

Milanowski, L., Pordzik, J., Janicki, P. K., Kaplon-Cieslicka, A., Rosiak, M., Peller, M., et al. (2017). New Single-Nucleotide Polymorphisms Associated with Differences in Platelet Reactivity and Their Influence on Survival in Patients with Type 2 Diabetes Treated with Acetylsalicylic Acid: an Observational Study. *Acta Diabetol.* 54 (4), 343–351. doi:10.1007/s00592-016-0945-y

Monrad, R. N., Errey, J. C., Barry, C. S., Iqbal, M., Meng, X., Iddon, L., et al. (2014). Dissecting the Reaction of Phase II Metabolites of Ibuprofen and Other NSAIDS with Human Plasma Protein. *Chem. Sci.* 5 (10), 3789–3794. doi:10.1039/c4sc01329h

Montalto, M., Gallo, A., Curigliano, V., D'Onofrio, F., Santoro, L., Covino, M., et al. (2010). Clinical Trial: the Effects of a Probiotic Mixture on Non-steroidal Anti-inflammatory Drug Enteropathy—a Randomized, Double-Blind, Cross-Over, Placebo-Controlled Study. *Aliment. Pharmacol. Ther.* 32 (2), 209–214. doi:10.1111/j.1365-2036.2010.04324.x

Montalto, M., Gallo, A., Gasbarrini, A., and Landolfi, R. (2013). NSAID Enteropathy: Could Probiotics Prevent it? *J. Gastroenterol.* 48 (6), 689–697. doi:10.1007/s00535-012-0648-2

Morham, S. G., Langenbach, R., Loftin, C. D., Tiano, H. F., Vouloumanos, N., Jennette, J. C., et al. (1995). Prostaglandin Synthase 2 Gene Disruption Causes Severe Renal Pathology in the Mouse. *Cell* 83 (3), 473–482. doi:10.1016/0092-8674(95)90125-6

Mortensen, B., Murphy, C., O'Grady, J., Lucey, M., Elsafi, G., Barry, L., et al. (2019). Bifidobacterium Breve Bif195 Protects against Small-Intestinal Damage Caused by Acetylsalicylic Acid in Healthy Volunteers. *Gastroenterology* 157 (3), 637–646. doi:10.1053/j.gastro.2019.05.008

Murray, F., O'Donovan, D G., Sheehan, K. M., and Murray, F. E. (1998). Prospective Evaluation of the Utilization of Aspirin and Non-steroidal Anti-inflammatory Drugs in Acute Medical Admissions. *Headache* 1, 3–7.

Musumba, C. O., Jorgensen, A., Sutton, L., Van Eker, D., Zhang, E., O'Hara, N., et al. (2013). CYP2C19*17 Gain-Of-Function Polymorphism Is Associated with Peptic Ulcer Disease. *Clin. Pharmacol. Ther.* 93 (2), 195–203. doi:10.1038/clpt.2012.215

Musumba, C., Pritchard, D. M., and Pirmohamed, M. (2009). Review Article: Cellular and Molecular Mechanisms of NSAID-Induced Peptic Ulcers. *Aliment. Pharmacol. Ther.* 30 (6), 517–531. doi:10.1111/j.1365-2036.2009.04086.x

Nicoletti, P., Carr, D. F., Barrett, S., McEvoy, L., Friedmann, P. S., Shear, N. H., et al. (2020). Beta-lactam-induced Immediate Hypersensitivity Reactions: A Genome-wide Association Study of a Deeply Phenotyped Cohort. *J. Allergy Clin. Immunol.* 147 (5), 1830.e15–1837.e15. doi:10.1016/j.jaci.2020.10.004

Niinuma, Y., Saito, T., Takahashi, M., Tsukada, C., Ito, M., Hirasawa, N., et al. (2014). Functional Characterization of 32 CYP2C9 Allelic Variants. *Pharmacogenomics J.* 14 (2), 107–114. doi:10.1038/tpj.2013.22

Niv, Y., Battler, A., Abuksis, G., Gal, E., Sapoznikov, B., and Vilkin, A. (2005). Endoscopy in Asymptomatic Minidose Aspirin Consumers. *Dig. Dis. Sci.* 50 (1), 78–80. doi:10.1007/s10620-005-1281-1

Olsen, A.-M. S., McGettigan, P., Gerds, T. A., Fosbøl, E. L., Olesen, J. B., Sindet-Pedersen, C., et al. (2011). Duration of Treatment with Nonsteroidal Anti-inflammatory Drugs and Impact on Risk of Death and Recurrent Myocardial Infarction in Patients with Prior Myocardial Infarction. *Circulation* 123 (20), 2226–2235.

Olsen, A.-M. S., McGettigan, P., Gerds, T. A., Fosbøl, E. L., Olesen, J. B., Sindet-Pedersen, C., et al. (2020). Risk of Gastrointestinal Bleeding Associated with Oral Anticoagulation and Non-steroidal Anti-inflammatory Drugs in Patients with Atrial Fibrillation: a Nationwide Study. *Eur. Heart J. Cardiovasc. Pharmacother.* 6 (5), 292–300. doi:10.1093/ehjcvp/pvz069

Onder, G., Pellicciotti, F., Gambassi, G., and Bernabei, R. (2004). NSAID-related Psychiatric Adverse Events. *Drugs* 64 (23), 2619–2627. doi:10.2165/00003495-200464230-00001

Ong, C. K. S., Lirk, P., Tan, C. H., and Seymour, R. A. (2007). An Evidence-Based Update on Nonsteroidal Anti-inflammatory Drugs. *Clinical Medicine and Research* 5 (1), 19–34. doi:10.3121/cmr.2007.698

Patrono, C., Coller, B., FitzGerald, G. A., Hirsh, J., and Roth, G. (2004). Platelet-Active Drugs: The Relationships Among Dose, Effectiveness, and Side Effects. *Chest* 126 (3), 234S–264S. doi:10.1378/chest.126.3_suppl.234s

Paulus, H. E. (1988). FDA Arthritis Advisory Committee Meeting: Serious Gastrointestinal Toxicity of Nonsteroidal Antiinflammatory Drugs; Drug-Containing Renal and Biliary Stones; Diclofenac and Carprofen Approved. *Arthritis Rheum.* 31 (11), 1450–1451. doi:10.1002/art.1780311118

Petrović, J., Pešić, V., and Lauschke, V. M. (2020). Frequencies of Clinically Important CYP2C19 and CYP2D6 Alleles Are Graded across Europe. *Eur. J. Hum. Genet.* 28 (1), 88–94. doi:10.1038/s41431-019-0480-8

PharmGKB, Pharmacogenomics Knowledgebase (PharmGKB) (2020). *Gene-specific Information Tables for CYP2C9.* https://www.pharmgkb.org/page/cyp2c9RefMaterials.

PharmVar (2017). *Pharmacogene Variation Consortium.* https://www.pharmvar.org/.

Pilotto, A., Franceschi, M., Leandro, G., and Di Mario, F. (2003). NSAID and Aspirin Use by the Elderly in General Practice. *Drugs and Aging* 20 (9), 701–710. doi:10.2165/00002512-200320090-00006

Pilotto, A., Franceschi, M., Vitale, D. F., Zaninelli, A., Masotti, G., and Rengo, F. (2005). Upper Gastrointestinal Symptoms and Therapies in Elderly Out-Patients, Users of Non-selective NSAIDs or Coxibs. *Aliment. Pharmacol. Ther.* 22 (2), 147–155. doi:10.1111/j.1365-2036.2005.02537.x

Pilotto, A., Seripa, D., Franceschi, M., Scarcelli, C., Colaizzo, D., Grandone, E., et al. (2007). Genetic Susceptibility to Nonsteroidal Anti-inflammatory Drug-Related Gastroduodenal Bleeding: Role of Cytochrome P450 2C9 Polymorphisms. *Gastroenterology* 133 (2), 465–471. doi:10.1053/j.gastro.2007.05.025

Pirmohamed, M., Aithal, G. P., Behr, E., Daly, A., and Roden, D. (2011). The Phenotype Standardization Project: Improving Pharmacogenetic Studies of Serious Adverse Drug Reactions. *Clin. Pharmacol. Ther.* 89 (6), 784–785. doi:10.1038/clpt.2011.30

Pirmohamed, M., Friedmann, P. S., Molokhia, M., Loke, Y. K., Smith, C., Phillips, E., et al. (2011). Phenotype Standardization for Immune-Mediated Drug-Induced Skin Injury. *Clin. Pharmacol. Ther.* 89 (6), 896–901. doi:10.1038/clpt.2011.79

Pirmohamed, M., James, S., Meakin, S., Green, C., Scott, A. K., Walley, T. J., et al. (2004). Adverse Drug Reactions as Cause of Admission to Hospital: Prospective Analysis of 18 820 Patients. *BMJ* 329 (7456), 15–19. doi:10.1136/bmj.329.7456.15

Plaza-Serón, M. d. C., García-Martín, E., Agúndez, J. A., and Ayuso, P. (2018). Hypersensitivity Reactions to Nonsteroidal Anti-inflammatory Drugs: an Update on Pharmacogenetics Studies. *Pharmacogenomics* 19 (13), 1069–1086. doi:10.2217/pgs-2018-0079

Quiralte, J., Sánchez-García, F., Torres, M.-J., Blanco, C., Castillo, R., Ortega, N., et al. (1999). Association of HLA-DR11 with the Anaphylactoid Reaction Caused by Nonsteroidal Anti-inflammatory Drugs. *J. Allergy Clin. Immunol.* 103 (4), 685–689. doi:10.1016/s0091-6749(99)70243-5

Rainsford, K. D. (2007). Anti-inflammatory Drugs in the 21st century. *Subcell Biochem.* 42, 3–27. doi:10.1007/1-4020-5688-5_1

Rainsford, K. D., and Whitehouse, M. W. (1980). Anti-inflammatory/anti-pyretic Salicylic Acid Esters with Low Gastric Ulcerogenic Activity. *Agents and Actions* 10 (5), 451–456. doi:10.1007/bf01968046

Rekatsina, M., Paladini, A., Cifone, M. G., Lombardi, F., Pergolizzi, J. V., and Varrassi, G. (2020). Influence of Microbiota on NSAID Enteropathy: A Systematic Review of Current Knowledge and the Role of Probiotics. *Adv. Ther.* 37 (5), 1933–1945. doi:10.1007/s12325-020-01338-6

Ricciotti, E., and FitzGerald, G. A. (2011). Prostaglandins and Inflammation. *Arterioscler. Thromb. Vasc. Biol.* 31 (5), 986–1000. doi:10.1161/atvbaha.110.207449

Rodrigues, A. D. (2005). Impact of Cyp2C9 Genotype on Pharmacokinetics: Are All Cyclooxygenase Inhibitors the Same?: Table 1. *Drug Metab. Dispos*. 33 (11), 1567–1575. doi:10.1124/dmd.105.006452

Rogers, M. A. M., and Aronoff, D. M. (2016). The Influence of Non-steroidal Anti-inflammatory Drugs on the Gut Microbiome. *Clin. Microbiol. Infect.* 22 (2), 178.e1–178.e9. doi:10.1016/j.cmi.2015.10.003

Rostom, A., Muir, K., Dubé, C., Jolicoeur, E., Boucher, M., Joyce, J., et al. (2007). Gastrointestinal Safety of Cyclooxygenase-2 Inhibitors: a Cochrane Collaboration Systematic Review. *Clin. Gastroenterol. Hepatol.* 5 (7), 818–828. doi:10.1016/j.cgh.2007.03.011

Russell, R. I. (2001). Non-steroidal Anti-inflammatory Drugs and Gastrointestinal Damage---problems and solutions. *Postgrad. Med. J.* 77 (904), 82–88. doi:10.1136/pmj.77.904.82

Samer, C. F., Lorenzini, K. I., Rollason, V., Daali, Y., and Desmeules, J. A. (2013). Applications of CYP450 testing in the clinical setting. *Mol. Diagn. Ther.* 17 (3), 165–184. doi:10.1007/s40291-013-0028-5

Sánchez-Diz, P., Estany-Gestal, A., Aguirre, C., Blanco, A., Carracedo, A., Ibáñez, L., et al. (2009). Prevalence of CYP2C9 polymorphisms in the south of Europe. *Pharmacogenomics J.* 9 (5), 306–310. doi:10.1038/tpj.2009.16

Sandler, R. S., Halabi, S., Baron, J. A., Budinger, S., Paskett, E., Keresztes, R., et al. (2003). A randomized trial of aspirin to prevent colorectal adenomas in patients with previous colorectal cancer. *N. Engl. J. Med.* 348 (10), 883–890. doi:10.1056/nejmoa021633

Sandson, N. B. (2015). "Uridine 5′-diphospho-glucuronoslytransferases (UGTs): Conjugating Cousins," in *A Case Approach to Perioperative Drug-Drug Interactions*. Editor C. Marcucci, et al. (New York, NY: Springer New York), 57–60. doi:10.1007/978-1-4614-7495-1_12

Sarganas, G., Buttery, A. K., Zhuang, W., Wolf, I. K., Grams, D., Rosario, A. S., et al. (2015). Prevalence, trends, patterns and associations of analgesic use in Germany. *BMC Pharmacol. Toxicol.* 16 (1), 28. doi:10.1186/s40360-015-0028-7

Schmassmann, A., Zoidl, G., Peskar, B. M., Waser, B., Schmassmann-Suhijar, D., Gebbers, J.-O., et al. (2006). Role of the different isoforms of cyclooxygenase and nitric oxide synthase during gastric ulcer healing in cyclooxygenase-1 and -2 knockout mice. *Am. J. Physiology-Gastrointestinal Liver Physiol.* 290 (4), G747–G756. doi:10.1152/ajpgi.00416.2005

Schwarz, U. I. (2003). Clinical relevance of genetic polymorphisms in the human CYP2C9 gene. *Eur. J. Clin. Invest.* 33 (Suppl. 2), 23–30. doi:10.1046/j.1365-2362.33.s2.6.x

Shaheen, N. J., Straus, W. L., and Sandler, R. S. (2002). Chemoprevention of gastrointestinal malignancies with nonsteroidal antiinflammatory drugs. *Cancer* 94 (4), 950–963. doi:10.1002/cncr.10333

Sharma, A., Buschmann, M. M., and Gilbert, J. A. (2019). Pharmacomicrobiomics: The Holy Grail to Variability in Drug Response? *Clin. Pharmacol. Ther.* 106 (2), 317–328. doi:10.1002/cpt.1437

Shen, Z.-H., Zhu, C.-X., Quan, Y.-S., Yang, Z.-Y., Wu, S., Luo, W.-W., et al. (2018). Relationship between intestinal microbiota and ulcerative colitis: Mechanisms and clinical application of probiotics and fecal microbiota transplantation. *Wjg* 24 (1), 5–14. doi:10.3748/wjg.v24.i1.5

Shiotani, A., Murao, T., Fujita, Y., Fujimura, Y., Sakakibara, T., Nishio, K., et al. (2013). Novel single nucleotide polymorphism markers for low dose aspirin-associated small bowel bleeding. *PloS one* 8 (12), e84244. doi:10.1371/journal.pone.0084244

Shiotani, A., Murao, T., Fujita, Y., Fujimura, Y., Sakakibara, T., Nishio, K., et al. (2014). Single nucleotide polymorphism markers for low-dose aspirin-associated peptic ulcer and ulcer bleeding. *J. Gastroenterol. Hepatol.* 29, 47–52. doi:10.1111/jgh.12770

Shiotani, A., Sakakibara, T., Nomura, M., Yamanaka, Y., Nishi, R., Imamura, H., et al. (2010). Aspirin-induced peptic ulcer and genetic polymorphisms. *J. Gastroenterol. Hepatol.* 25, S31–S34. doi:10.1111/j.1440-1746.2009.06212.x

Sigthorsson, G., Simpson, R. J., Walley, M., Anthony, A., Foster, R., Hotz–Behoftsitz, C., et al. (2002). COX-1 and 2, intestinal integrity, and pathogenesis of nonsteroidal anti-inflammatory drug enteropathy in mice. *Gastroenterology* 122 (7), 1913–1923. doi:10.1053/gast.2002.33647

Silverstein, F. E., Graham, D. Y., Senior, J. R., Davies, H. W., Struthers, B. J., and Bittman, R. M. (1995). Misoprostol Reduces Serious Gastrointestinal Complications in Patients with Rheumatoid Arthritis Receiving Nonsteroidal Anti-Inflammatory Drugs. *Ann. Intern. Med.* 123 (4), 241–249. doi:10.7326/0003-4819-123-4-199508150-00001

Silverstein, F. E., Faich, G., Goldstein, J. L., Simon, L. S., Pincus, T., Whelton, A., et al. (2000). Gastrointestinal Toxicity With Celecoxib vs Nonsteroidal Anti-inflammatory Drugs for Osteoarthritis and Rheumatoid Arthritis. *JAMA* 284 (10), 1247–1255. doi:10.1001/jama.284.10.1247

Sim, S. C., Kacevska, M., and Ingelman-Sundberg, M. (2013). Pharmacogenomics of drug-metabolizing enzymes: a recent update on clinical implications and endogenous effects. *Pharmacogenomics J.* 13 (1), 1–11. doi:10.1038/tpj.2012.45

Singh, G, and Triadafilopoulos, G (1999). Epidemiology of NSAID induced gastrointestinal complications. *J. Rheumatol. Suppl.* 56, 18–24.

Singh, G., Ramey, D. R., Morfeld, D., Shi, H., Hatoum, H. T., and Fries, J. F. (1996). Gastrointestinal tract complications of nonsteroidal anti-inflammatory drug treatment in rheumatoid arthritis. A prospective observational cohort study. *Arch. Intern. Med.* 156 (14), 1530–1536. doi:10.1001/archinte.156.14.1530

Sorensen, H. T., Mellemkjaer, L., Blot, W. J., Nielsen, G. L., Steffensen, F. H., McLaughlin, J. K., et al. (2000). Risk of upper gastrointestinal bleeding associated with use of low-dose aspirin. *Am. J. Gastroenterol.* 95 (9), 2218–2224. doi:10.1111/j.1572-0241.2000.02248.x

Sostres, C., Gargallo, C. J., Arroyo, M. T., and Lanas, A. (2010). Adverse effects of non-steroidal anti-inflammatory drugs (NSAIDs, aspirin and coxibs) on upper gastrointestinal tract. *Best Pract. Res. Clin. Gastroenterol.* 24 (2), 121–132. doi:10.1016/j.bpg.2009.11.005

Sostres, C., Gargallo, C. J., and Lanas, A. (2013). Nonsteroidal anti-inflammatory drugs and upper and lower gastrointestinal mucosal damage. *Arthritis Res. Ther.* 15 (Suppl. 3Suppl 3), S3. doi:10.1186/ar4175

Sostres, C., and Lanas, A. (2011). Should Prophylactic Low-dose Aspirin Therapy be Continued in Peptic Ulcer Bleeding? *Drugs* 71 (1), 1–10. doi:10.2165/11585320-000000000-00000

Speed, W. C., Kang, S. P., Tuck, D. P., Harris, L. N., and Kidd, K. K. (2009). Global variation in CYP2C8-CYP2C9 functional haplotypes. *Pharmacogenomics J.* 9 (4), 283–290. doi:10.1038/tpj.2009.10

Stamer, U. M., Zhang, L., and Stüber, F. (2010). Personalized therapy in pain management: where Do we stand? *Pharmacogenomics* 11 (6), 843–864. doi:10.2217/pgs.10.47

Starodub, O. T., Demitrack, E. S., Baumgartner, H. K., and Montrose, M. H. (2008). Disruption of the Cox-1 gene slows repair of microscopic lesions in the mouse gastric epithelium. *Am. J. Physiology-Cell Physiol.* 294 (1), C223–C232. doi:10.1152/ajpcell.00395.2006

Stillhart, C., Vučićević, K., Augustijns, P., Basit, A. W., Batchelor, H., Flanagan, T. R., et al. (2020). Impact of gastrointestinal physiology on drug absorption in special populations--An UNGAP review. *Eur. J. Pharm. Sci.* 147, 105280. doi:10.1016/j.ejps.2020.105280

Stosic, R., Dunagan, F., Palmer, H., Fowler, T., and Adams, I. (2011). Responsible self-medication: perceived risks and benefits of over-the-counter analgesic use. *Int. J. Pharm. Pract.* 19 (4), 236–245. doi:10.1111/j.2042-7174.2011.00097.x

Suzuki, T., Masui, A., Nakamura, J., Shiozawa, H., Aoki, J., Nakae, H., et al. (2017). Yogurt Containing Lactobacillus gasseri Mitigates Aspirin-Induced Small Bowel Injuries: A Prospective, Randomized, Double-Blind, Placebo-Controlled Trial. *Digestion* 95 (1), 49–54. doi:10.1159/000452361

Syer, S. D., and Wallace, J. L. (2014). Environmental and NSAID-Enteropathy: Dysbiosis as a Common Factor. *Curr. Gastroenterol. Rep.* 16 (3), 377. doi:10.1007/s11894-014-0377-1

Szeto, C.-C., Sugano, K., Wang, J.-G., Fujimoto, K., Whittle, S., Modi, G. K., et al. (2020). Non-steroidal anti-inflammatory drug (NSAID) therapy in patients with hypertension, cardiovascular, renal or gastrointestinal comorbidities: joint APAGE/APLAR/APSDE/APSH/APSN/PoA recommendations. *Gut* 69 (4), 617–629. doi:10.1136/gutjnl-2019-319300

Tanaka, A., Araki, H., Hase, S., Komoike, Y., and Takeuchi, K. (2002). Up-regulation of COX-2 by inhibition of COX-1 in the rat: a key to NSAID-induced gastric injury. *Aliment. Pharmacol. Ther.* 16 (Suppl. 2), 90–101. doi:10.1046/j.1365-2036.16.s2.22.x

Tanaka, A., Araki, H., Komoike, Y., Hase, S., and Takeuchi, K. (2001). Inhibition of Both COX-1 and COX-2 is required for development of gastric damage in response to nonsteroidal antiinflammatory drugs. *J. Physiol. Paris* 95 (1-6), 21–27. doi:10.1016/s0928-4257(01)00005-5

Theken, K. N., Lee, C. R., Gong, L., Caudle, K. E., Formea, C. M., Gaedigk, A., et al. (2020). Clinical Pharmacogenetics Implementation Consortium Guideline (CPIC) for CYP2C9 and Nonsteroidal Anti-Inflammatory Drugs. *Clin. Pharmacol. Ther.* 108 (2), 191–200. doi:10.1002/cpt.1830

Theken, K. N., Lee, C. R., Gong, L., Caudle, K. E., Formea, C. M., Gaedigk, A., et al. (2020). "Supplemental Material Clinical Pharmacogenetics Implementation Consortium (CPIC) Guideline for CYP2C9 and Nonsteroidal Anti-inflammatory Drugs. *CPIC GUIDELINE.* Editors.

Thiele, I., Clancy, C. M., Heinken, A., and Fleming, R. M. T. (2017). Quantitative systems pharmacology and the personalized drug-microbiota-diet axis. *Curr. Opin. Syst. Biol.* 4, 43–52. doi:10.1016/j.coisb.2017.06.001

To, K. F., Chan, F. K. L., Cheng, A. S. L., Lee, T. L., Ng, Y. P., and Sung, J. J. Y. (2001). Up-regulation of cyclooxygenase-1 and -2 in human gastric ulcer. *Aliment. Pharmacol. Ther.* 15 (1), 25–34. doi:10.1046/j.1365-2036.2001.00889.x

Turner, R. M., de Koning, E. M., Fontana1, V., Thompson1, A., and Pirmohamed, M. (2020). Multimorbidity, polypharmacy, and drug-drug-gene interactions following a non-ST elevation acute coronary syndrome: analysis of a multicentre observational study. *BMC Med.* 18 (1), 367. doi:10.1186/s12916-020-01827-z

Uijen, A. A., and van de Lisdonk, E. H. (2008). Multimorbidity in primary care: prevalence and trend over the last 20 years. *Eur. J. Gen. Pract.* 14 (Suppl. 1), 28–32. doi:10.1080/13814780802436093

Ulisse, S., Gionchetti, P., D'Alo, S., Russo, F. P., Pesce, I., Ricci, G., et al. (2001). Expression of cytokines, inducible nitric oxide synthase, and matrix metalloproteinases in pouchitis: effects of probiotic treatment. *Am. J. Gastroenterol.* 96 (9), 2691–2699. doi:10.1111/j.1572-0241.2001.04139.x

Ulrich, C. M., Bigler, J., and Potter, J. D. (2006). Non-steroidal anti-inflammatory drugs for cancer prevention: promise, perils and pharmacogenetics. *Nat. Rev. Cancer* 6 (2), 130–140. doi:10.1038/nrc1801

Van Booven, D., Marsh, S., McLeod, H., Carrillo, M. W., Sangkuhl, K., Klein, T. E., et al. (2010). Cytochrome P450 2C9-CYP2C9. *Pharmacogenetics and genomics* 20 (4), 277–281. doi:10.1097/fpc.0b013e3283349e84

Van Hecken, A, Schwartz, JI, Depré, M, De Lepeleire, I, Dallob, A, Tanaka, W, et al. (2000). Comparative inhibitory activity of rofecoxib, meloxicam, diclofenac, ibuprofen, and naproxen on COX-2 versus COX-1 in healthy volunteers. *J. Clin. Pharmacol.* 40 (10), 1109–20.

van Oijen, M. G. H., Huybers, S., Peters, W. H. M., Drenth, J. P. H., Laheij, R. J. F., Verheugt, F. W. A., et al. (2005). Polymorphisms in genes encoding acetylsalicylic acid metabolizing enzymes are unrelated to upper gastrointestinal health in cardiovascular patients on acetylsalicylic acid. *Br. J. Clin. Pharmacol.* 60 (6), 623–628. doi:10.1111/j.1365-2125.2005.02495.x

van Oijen, M. G. H., Laheij, R. J. F., Koetsier, M., de Kleine, E., te Morsche, R. H. M., van Kerkhoven, L. A. S., et al. (2006). Effect of a specific cyclooxygenase-gene polymorphism (A-842G/C50T) on the occurrence of peptic ulcer hemorrhage. *Dig. Dis. Sci.* 51 (12), 2348–2352. doi:10.1007/s10620-006-9475-8

Vane, J. R. (1971). Inhibition of prostaglandin synthesis as a mechanism of action for aspirin-like drugs. *Nat. New Biol.* 231 (25), 232–235. doi:10.1038/newbio231232a0

Visser, L., Vanschaik, R., Vanvliet, M., Trienekens, P., Desmet, P., Vulto, A., et al. (2005). Allelic variants of cytochrome P450 2C9 modify the interaction between nonsteroidal anti-inflammatory drugs and coumarin anticoagulants. *Clin. Pharmacol. Ther.* 77 (6), 479–485. doi:10.1016/j.clpt.2005.02.009

Vonkeman, H. E., van de Laar, M. A. F. J., van der Palen, J., Brouwers, J. R. B. J., and Vermes, I. (2006). Allele variants of the cytochrome P450 2C9 genotype in white subjects from The Netherlands with serious gastroduodenal ulcers attributable to the use of NSAIDs. *Clin. Ther.* 28 (10), 1670–1676. doi:10.1016/j.clinthera.2006.10.019

Wadelius, M., Chen, L. Y., Lindh, J. D., Eriksson, N., Ghori, M. J. R., Bumpstead, S., et al. (2009). The largest prospective warfarin-treated cohort supports genetic forecasting. *Blood* 113 (4), 784–792. doi:10.1182/blood-2008-04-149070

Wallace, J. L., McKnight, W., Reuter, B. K., and Vergnolle, N. (2000). NSAID-induced gastric damage in rats: Requirement for inhibition of Both cyclooxygenase 1 and 2. *Gastroenterology* 119 (3), 706–714. doi:10.1053/gast.2000.16510

Wallace, J. L. (1992). Prostaglandins, NSAIDs, and cytoprotection. *Gastroenterol. Clin. North America* 21 (3), 631–641. doi:10.1016/s0889-8553(21)00052-2

Wallace, J. L. (2008). Prostaglandins, NSAIDs, and gastric mucosal protection: why doesn't the stomach digest itself? *Physiol. Rev.* 88 (4), 1547–1565. doi:10.1152/physrev.00004.2008

Walt, R., Logan, R., Katschinski, B., Ashley, J., and Langman, M. (1986). Rising frequency of ulcer perforation in elderly people in the United Kingdom. *The Lancet* 327 (8479), 489–492. doi:10.1016/s0140-6736(86)92940-5

Wang, L., McLeod, H. L., and Weinshilboum, R. M. (2011). Genomics and drug response. *N. Engl. J. Med.* 364 (12), 1144–1153. doi:10.1056/nejmra1010600

Weil, J., Langmana, M. J. S., Wainwrighta, P., Lawsone, D. H., Rawlinsb, M., Loganc, R. F. A., et al. (2000). Peptic ulcer bleeding: accessory risk factors and interactions with non-steroidal anti-inflammatory drugs. *Gut* 46 (1), 27–31. doi:10.1136/gut.46.1.27

Wilcox, CM, Cryer, B, and Triadafilopoulos, G (2005). Patterns of use and public perception of over-the-counter pain relievers: focus on nonsteroidal antiinflammatory drugs. *J. Rheumatol.* 32 (11), 2218–24.

Wolfe, M. M., Lichtenstein, D. R., and Singh, G. (1999). Gastrointestinal toxicity of nonsteroidal antiinflammatory drugs. *N. Engl. J. Med.* 340 (24), 1888–1899. doi:10.1056/nejm199906173402407

Wongrakpanich, S., Wongrakpanich, A., Melhado, K., and Rangaswami, J. (2018). A Comprehensive Review of Non-Steroidal Anti-Inflammatory Drug Use in The Elderly. *AandD* 9 (1), 143–150. doi:10.14336/ad.2017.0306

Wu, H., Esteve, E., Tremaroli, V., Khan, M. T., Caesar, R., Mannerås-Holm, L., et al. (2017). Metformin alters the gut microbiome of individuals with treatment-naive type 2 diabetes, contributing to the therapeutic effects of the drug. *Nat. Med.* 23 (7), 850–858. doi:10.1038/nm.4345

Wu, Y., Hu, Y., You, P., Chi, Y.-J., Zhou, J.-H., Zhang, Y.-Y., et al. (2016). Study of Clinical and Genetic Risk Factors for Aspirin-induced Gastric Mucosal Injury. *Chin. Med. J.* 129 (2), 174–180. doi:10.4103/0366-6999.173480

Xie, H.-G., Prasad, H. C., Kim, R. B., and Stein, C. M. (2002). CYP2C9 allelic variants: ethnic distribution and functional significance. *Adv. Drug Deliv. Rev.* 54 (10), 1257–1270. doi:10.1016/s0169-409x(02)00076-5

Yip, L. Y., and Chan, E. C. Y. (2015). Investigation of Host-Gut Microbiota Modulation of Therapeutic Outcome. *Drug Metab. Dispos.* 43 (10), 1619–1631. doi:10.1124/dmd.115.063750

Zhou, Y., Boudreau, D. M., and Freedman, A. N. (2014). Trends in the use of aspirin and nonsteroidal anti-inflammatory drugs in the general U.S. population. *Pharmacoepidemiol. Drug Saf.* 23 (1), 43–50. doi:10.1002/pds.3463

Zimmermann, M., Zimmermann-Kogadeeva, M., Wegmann, R., and Goodman, A. L. (2019). Mapping human microbiome drug metabolism by gut bacteria and their genes. *Nature* 570 (7762), 462–467. doi:10.1038/s41586-019-1291-3

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 McEvoy, Carr and Pirmohamed. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Human Genetics
https://doi.org/10.1007/s00439-020-02143-5

ORIGINAL INVESTIGATION

A genomics approach to females with infertility and recurrent pregnancy loss

Sateesh Maddirevula[1] · **Khalid Awartani**[2] · **Serdar Coskun**[3] · **Latifa F. AlNaim**[2] · **Niema Ibrahim**[1] · **Firdous Abdulwahab**[1] · **Mais Hashem**[1] · **Saad Alhassan**[2] · **Fowzan S. Alkuraya**[1,4]

Received: 29 December 2019 / Accepted: 19 February 2020
© Springer-Verlag GmbH Germany, part of Springer Nature 2020

Abstract

Infertility affects 10% of reproductive-age women and is extremely heterogeneous in etiology. The genetic contribution to female infertility is incompletely understood, and involves chromosomal and single-gene defects. Our aim in this study is to decipher single-gene causes in infertile women in whom endocrinological, anatomical, and chromosomal causes have been excluded. Our cohort comprises women with recurrent pregnancy loss and no offspring from spontaneous pregnancies (RPL, $n=61$) and those who never achieved clinical pregnancy and were referred for in vitro fertilization [primary infertility (PI), $n=14$]. Whole-exome sequencing revealed candidate variants in 14, which represents 43% of those with PI and 13% of those with RPL. These include variants in previously established female infertility-related genes (*TLE6*, *NLRP7*, *FSHR*, and *ZP1*) as well as genes with only tentative links in the literature (*NLRP5*). Candidate variants in genes linked to primary ciliary dyskinesia (*DNAH11* and *CCNO*) were identified in individuals with and without systemic features of the disease. We also identified variants in genes not previously linked to female infertility. These include one homozygous variant each in *CCDC68*, *CBX3*, *CENPH*, *PABPC1L*, *PIF1*, *PLK1*, and *REXO4*, which we propose as candidate genes for infertility based on their established biology or compatible animal models. Our study expands the contribution of single genes to the etiology of PI and RPL, improves the precision of disease classification at the molecular level, and offers the potential for future treatment and development of human genetics-inspired fertility regulators.

Keywords Primary ciliary dyskinesia · Primary infertility · Preimplantation embryonic lethality · Meiotic arrest · Fertilization failure · Recurrent pregnancy loss

Introduction

Infertility is a relatively common reproductive phenotype that affects 10–15% of couples worldwide. In the US alone, it is estimated that 7.7 million women will be diagnosed with infertility in 2025 (Stephen and Chandra 1998). Primary infertility refers to the inability to achieve clinical pregnancy after 1 year of regular and unprotected sexual intercourse, and this can be due to male and female factors (Zegers-Hochschild et al. 2017). Infertility in female is extremely heterogeneous, which reflects the complex interaction of numerous developmental, hormonal, environmental, and genetic factors (Beke 2019). The usual workup of infertile women includes a careful review of systemic diseases, hormonal assessment, anatomical scans, and karyotyping. Such a workup has a limited yield and leaves many women with the uncomfortable label of idiopathic infertility.

Sateesh Maddirevula, Khalid Awartani, and Serdar Coskun have contributed equally.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00439-020-02143-5) contains supplementary material, which is available to authorized users.

✉ Fowzan S. Alkuraya
falkuraya@kfshrc.edu.sa

[1] Department of Genetics, King Faisal Specialist Hospital and Research Center, Riyadh, Saudi Arabia

[2] Department of Obstetrics and Gynecology, King Faisal Specialist Hospital and Research Center, Riyadh, Saudi Arabia

[3] Department of Pathology and Laboratory Medicine, King Faisal Specialist Hospital and Research Center and College of Medicine, Alfaisal University, Riyadh, Saudi Arabia

[4] Department of Anatomy and Cell Biology, College of Medicine, Alfaisal University, Riyadh, Saudi Arabia

Published online: 14 March 2020

The murine animal model has served as a very helpful mammalian proxy to decipher the molecular basis of female fertility through the systematic inactivation of single genes and investigation of the precise consequences of their deficiencies (Greenhouse et al. 1998). Thousands of such genes have been identified and their roles span the entire spectrum from embryonic ovarian development to maternal factors influencing early embryonic development. Unfortunately, only a very small fraction of these genes have been linked to female infertility in humans, mostly due to the fact that exome and genome sequencing has not yet made inroads in infertility clinics. The ~ 200 genes identified to date should serve as an impetus to continue this line of research, since they demonstrated both clinical and scientific utility. For example, we now know that IVF is futile and should not be pursued in women with biallelic mutations of genes involved in early zygote cleavage such as *TLE6* (Alazami et al. 2015). Furthermore, the identification of infertility-causing variants in the ZP genes sets the stage for the development of novel contraceptives targeting the zona pellucida (Nishimura et al. 2019).

Dominant, recessive, and X-linked variants have all been described in the previous discoveries of female infertility-related genes (Group 2008). The highly consanguineous population of Saudi Arabia is known to be enriched for recessive disorders, and the population structure lends itself to positional mapping, especially when combined with WES (Alkuraya 2014). Indeed, we have previously exploited these advantages to identify *TLE6, PADI6,* and *PATL2* as important determinants in female infertility (Alazami et al. 2015; Maddirevula et al. 2017a, b). Here, we extend this work to a much larger cohort and show a corresponding boost in the identification of variants in both known as well as novel disease genes.

Materials and methods

Human subjects

We recruited cases that presented to our infertility center with either (a) primary infertility (PI) or (b) recurrent pregnancy loss (RPL, ≥ 3 losses) with no offspring from spontaneous pregnancies. Cases with PI were recruited if they met any of the following criteria: (1) positive family history, (2) recurrent low oocyte maturity (RLOM), (3) repeated fertilization failure, (4) recurrent implantation failure, and (5) empty follicle syndrome (EFS). Females with polycystic ovary syndrome (PCOS) were only eligible if there was a documented failure to achieve pregnancy after several ovulation inductions with gonadotropin and intra-uterine insemination (IUI). On the other hand, all cases with recurrent pregnancy loss between April 2018 and Sept 2019 were included. We only included cases whose male partners had normal sperm count according to WHO 5th edition (World Health Organization 2010). *MTHFR* and *F2* genes were screened by commercial genotyping assays and the outcome, along with all other clinical laboratory measurements, is documented in Table S1. We also excluded cases with chromosomal, endocrinological or anatomical causes. We defined pregnancy loss ≥ 10 weeks' size with a fetus (≥ 33 mm) on ultrasound as fetal loss, intra-uterine pregnancy loss with an embryo without cardiac activity on ultrasound as embryonic loss, and pregnancy demise based on decreasing serum or urinary β-hCG levels without an ultrasound evaluation as biochemical pregnancy loss (Table S1). Informed consent was obtained from all participants and blood samples were collected in EDTA tubes for genomic DNA extraction and subsequent analysis. This study was approved by the local IRB (KFSHRC RAC# 2140016 and 2121053).

Positional mapping, exome sequencing (ES), and variant prioritization

ES was performed on the index patient of each family as described before (Monies et al. 2017). Given the high degree of consanguinity in the study population, we prioritized variants that are within the autozygome. To combine positional mapping with ES, we employed the method previously described to call the autozygome using ES variants (Monies et al. 2017). We first searched for candidate variants in genes with listed Online Mendelian Inheritance in Man (OMIM) phenotypes. If none were identified, we searched for candidate variants in genes with tentative links to infertility based on their biological roles or corresponding animal models. We followed the ACMG guidelines for variant classification, and only considered variants that are absent or very rare (MAF < 0.00001) in gnomAD and 2379 ethnically matched exomes and had a CAD score of > 15.

Results

Clinical phenotypes

Our cohort comprises women with RPL with no offspring from spontaneous pregnancies ($n = 61$) and those who never achieved clinical pregnancy and were referred for IVF (PI, $n = 14$). Details of their phenotypes are shown in Table S1.

Identification of single-gene causes of female with infertility and RPL

Variants that are potentially relevant to the various phenotypes of female infertility were identified in 18.66% of cases

($n=14$), including eight variants that are pathogenic/likely pathogenic according to the ACMG classification (Fig. 1; Table 1). A breakdown by phenotype revealed that we identified variants in 43% ($n=6$) of PI and 13% ($n=8$) of RPL patients. These variants are further described below.

1. Genes with established links to PI or RPL:
    - (i.) *TLE6*: We identified a homozygous truncating variant NM_024760.2:c.436_437del;p.(Lys146Glufs*51) in an individual with history of four biochemical pregnancies. Only fragmented embryos could be seen following IVF (Fig. 2a; Supplementary Video 1). This is a gene already linked to preimplantation embryonic lethality (Alazami et al. 2015).
    - (ii.) *NLRP7*: This gene is already linked to recurrent hydatidiform mole. In an individual with recurrent molar pregnancy who sought treatment, we observed no pregnancy following three intra-uterine insemination attempts. From two IVF cycles, 12 oocytes were retrieved. Six were intra-cytoplasmic sperm injected, but none was fertilized. The other six were inseminated by IVF and displayed three pronuclei, which predict the development of hydatidiform mole (Fig. 2b). This individual was found to have a homozygous truncating variant in *NLRP7*: NM_206828.3:c.1857_1858del;p.(Lys619Asnfs*18).
    - (iii.) *ZP1*: We encountered one individual with empty follicle syndrome. Upon combined stimulation with rFSH and hMG, she developed 22 follicles (> 10 mm); however, following ovum pickup, no oocytes were found (Fig. 2c). WES revealed a homozygous truncating variant in *ZP1*: NM_207341.3:c.1168delA;p.(Ile390Phefs*10). This gene is already linked to oocyte maturation defect (OMIM: 615774).
    - (iv.) *FSHR*: This gene is already linked to ovarian failure. An individual who is homozygous for the likely deleterious variant

Fig. 1 Summary of the study findings

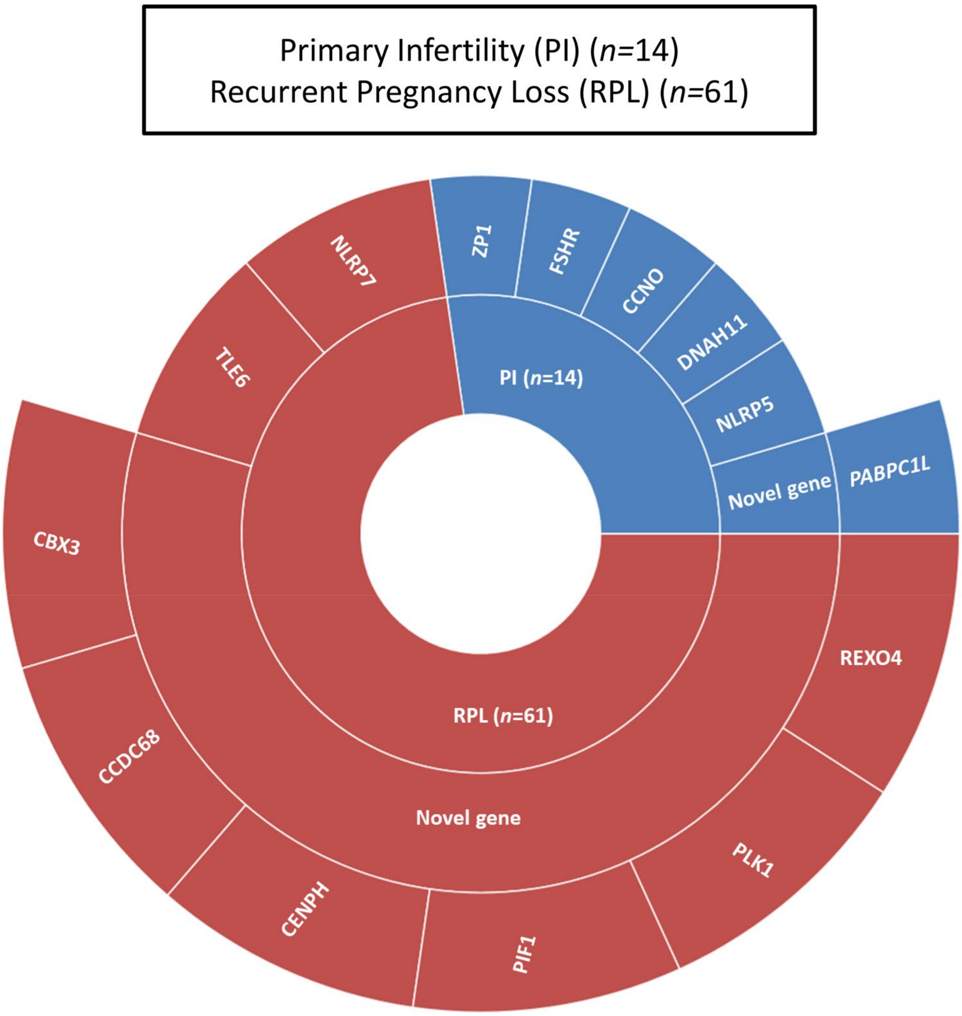


Table 1 List of homozygous variants identified in the study cohort with RPL and PI

Case ID	Pheno-type	Gene	Variant	Novelty of the gene	OMIM	CADD	TraP	ACMG classification	Chromosomal variant (hg38)	gnomAD (v3)	CliniVar
19DG2709	PI	*NLRP5*	NM_153447.4:c.2274_2275del;p.(Trp759Aspfs*4)	Known gene	NA	NA	NA	Likely pathogenic (PVS1;PM2)	NC_000019.10:g.56028507_56028508del	Absent	SCV001161790
19DG2698	PI	*PABPC1L*	NM_001372179.1:c.77T>A;p.(Met26Lys)	Novel gene	NA	23.1	NA	Uncertain significance	NC_000020.11:g.44910220T>A	0.000006977 (1 het)	SCV001161791
19DG2714	PI	*DNAH11*	NM_001277115.1:c.9685C>T;p.(Arg3229Trp)	Known gene	Ciliary dyskinesia, primary, 7, with or without situs inversus	22.8	NA	Uncertain significance (PM1;PM2;PP3)	NC_000007.14:g.21786711C>T	0.00001396 (2 het)	SCV001161792
19DG2718	PI	*CCNO*	NM_021147.3:c.964delC;p.(Leu322Cysfs*5)	Known gene	Ciliary dyskinesia, primary, 29	NA	NA	Likely pathogenic (PVS1;PM2)	NC_000005.10:g.55231464delG	Absent	SCV001161793
19DG2720	PI	*FSHR*	NM_000145.3:c.1121T>A;p.(Ile374Asn)	Known gene	Ovarian dysgenesis 1/ovarian response to FSH stimulation	25.8	NA	Likely pathogenic (PM2;PP1;PP3)	NC_000002.12:g.48963700A>T	Absent	SCV001161794
19DG2721	PI	*ZP1*	NM_207341.3:c.1168delA;p.(Ile390Phefs*10)	Known gene	Oocyte maturation defect 1	NA	NA	Likely pathogenic (PVS1;PM2)	NC_000011.10:g.60873217delA	Absent	SCV001161795
19DG2708	RPL	*PIF1*	NM_025049.3:c.506C>T;p.(Thr169Ile)	Novel gene	NA	15.3	NA	Uncertain significance	NC_000015.10:g.64823830G>A	Absent	SCV001161796
19DG2717	RPL	*CCDC68*	NM_001143829.1:c.775dupA;p.(Ser259Lysfs*12)	Novel gene	NA	NA	NA	Uncertain significance	NC_000018.10:g.54919285dupT	Absent	SCV001161797
19DG2710	RPL	*TLE6*	NM_024760.2:c.436_437del;p.(Lys146Glufs*51)	Known gene	Preimplantation embryonic lethality	NA	NA	Likely pathogenic (PVS1;PM2)	NC_000019.10:g.2989125_2989126del	Absent	SCV001161798
19DG0738	RPL	*CBX3*	NM_007276.4:c.330+3A>G	Novel gene	NA	NA	0.46	Uncertain significance	NC_000007.14:g.26208558A>G	0.000008069 (1 het)	SCV001161799
19DG2713	RPL	*CENPH*	NM_022909.3:c.635T>C;p.(Ile212Thr)	Novel gene	NA	24	NA	Uncertain significance	NC_000005.10:g.69208343T>C	Absent	SCV001161800
19DG2695	RPL	*REXO4*	NM_001279350.1:c.253C>T;p.(Gln85*)	Novel gene	NA	35	NA	Uncertain significance	NC_000009.12:g.133412830G>A	Absent	SCV001161801
19DG2719	RPL	*NLRP7*	NM_206828.3:c.1857_1858del;p.(Lys619Asnfs*18)	Known gene	Hydatidiform mole, recurrent, 1	NA	NA	Likely pathogenic (PVS1;PM2)	NC_000019.10:g.54938961_54938962del	Absent	SCV001161802
19DG0494	RPL	*PLK1*	NM_005030.3:c.128T>C;p.(Val43Ala)	Novel gene	NA	23.1	NA	Uncertain significance	NC_000016.10:g.23679060T>C	Absent	SCV001161803

CADD combined annotation-dependent depletion, *NA* not applicable, *RPL* recurrent pregnancy loss, *PI* primary infertility, *TraP* transcript-inferred pathogenicity score

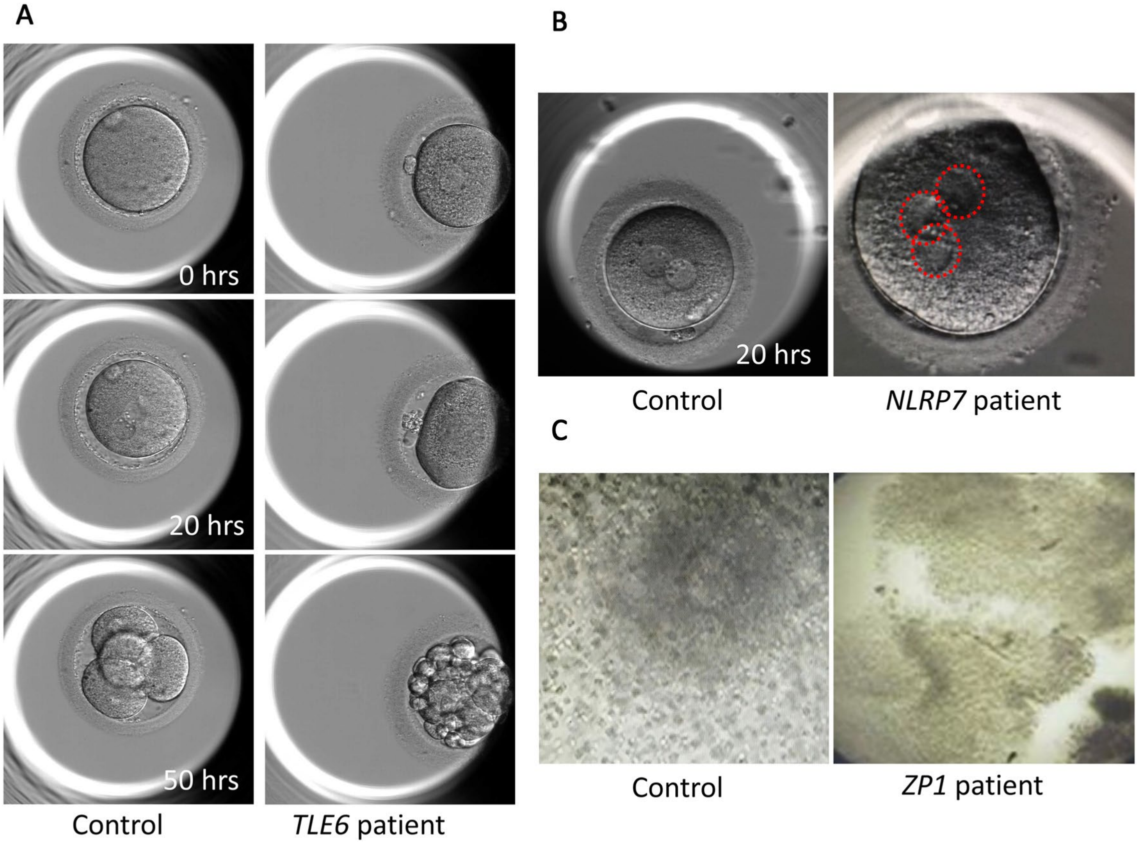


Fig. 2 **a** Oocytes form a patient with *TLE6* mutation. Control and patient oocytes at 0 h after sperm injection are shown. At 20 h, patient oocyte displays very faint two pronuclei in contrast to control. At 50 h, patient oocyte displays four-cell blastomere with heavy fragmentation in contrast to control. **b** Oocyte from a patient with *NLRP7* mutation exhibiting three pronuclei at 20 h in contrast to control. **c** A patient with a mutation in *ZP1* displaying cumulus-like cells with no oocyte in contrast to control

NM_000145.3:c.1121T>A;p.(Ile374Asn) presented with 22 follicles (<5 mm) at baseline ultrasound scan. Ovarian stimulation was continued for 25 days; however, no follicular growth was observed (all remained <5 mm in size). Her two infertile sisters reported a similar history of poor follicular growth when treated in other centers and shared the same homozygous variant.

2. Genes with tentative links to PI or RPL:

   (i.) *NLRP5*: A recent report described *NLRP2* and *NLRP5* variants in individuals with early embryonic arrest (Mu et al. 2019). We identified a PI individual who had two failed IVF cycles. In the first cycle, 28 oocytes were obtained, 14 of which were assigned for intracytoplasmic sperm injection (ICSI). Eight of these were matured; however, upon ICSI, none were fertilized. The other 14 oocytes were inseminated by IVF, but there was also no sign of fertilization. In the second cycle, 18 oocytes were obtained, 6 of which were matured; however, none were fertilized. ES revealed a homozygous truncating variant *NLRP5* (NM_153447.4:c.2274_2275del;p.(Trp759Aspfs*4), thus lending further support to the previously published preliminary association.

   (ii.) PCD genes: Although male infertility is a well-established aspect of the primary ciliary dyskinesia (PCD) phenotype, female infertility in PCD is much less clear (Raidt et al. 2015). In our cohort, we identified two individuals with homozygous likely deleterious variants in PCD genes *DNAH11* and *CCNO*. Only the individual with a homozygous truncating variant in *CCNO* presented with a classical PCD phenotype, whereas that with the *DNAH11* variant presented with PI only (Table S1). Interestingly, IVF was pursued in both patients. The patient with *CCNO* variant had one fresh and two frozen embryo transfer cycles, and one of the latter resulted in an ectopic pregnancy. The patient with *DNAH11* had three cycles of IVF and had one clinical pregnancy that, unfortunately, ended in a spontaneous loss during first trimester.

3. Genes not known to cause PI or RPL in humans:

    (i.) *CCDC68*: The encoded centrosome-binding protein is necessary for the hierarchical assembly of subdistal appendage (SDA), which in turn is required for the anchoring of microtubules in interphase cells (Huang et al. 2017). A 35 year old individual who had two recurrent miscarriages and had a baby through IVF was found on WES to have a homozygous truncating variant NM_001143829.1:c.775dupA;p.(Ser259Lysfs*12). Interestingly, this same variant was also found in our lab's collection in the homozygous state in another individual who had a history of recurrent miscarriages. However, she was not eligible for inclusion in this study, because she eventually had a term pregnancy.

    (ii.) *CBX3*: The encoded protein HP1 gamma plays an important role in regulating the cell cycle in primordial germ cells (PGCs), such that *Cbx3*$^{-/-}$ mouse embryos are depleted in PGCs with resulting sterility phenotype in adult males and females (Abe et al. 2011; Brown et al. 2010). We identified a homozygous potentially splicing variant *CBX3* (NM_007276.4:c.330+3A>G) in an individual with recurrent pregnancy loss and history of two failed IVF cycles.

    (iii.) *PABPC1L*: This poly(A)-binding protein is required for oocyte maturation and its deficiency causes female infertility in mouse (Guzeloglu-Kayisli et al. 2012). We identified a homozygous likely deleterious variant in *PABPC1L* (NM_001372179.1:c.77T>A;p.(Met26Lys)) in a woman whose infertility phenotype resembles that of *Pabpc1l*$^{-/-}$ mouse. During her IVF cycles, 18 oocytes were retrieved and subjected to IVF and ICSI. Nine oocytes were assigned to ICSI, but eight were at germinal vesicle stage and only one showed polar body and failed to fertilize following ICSI. Similarly, nine oocytes were assigned to IVF, and only two showed polar body on the next day without any sign of fertilization. The remaining oocytes were at germinal vesicle stage.

    (iv.) *CENPH*: Complete deficiency of this kinetochore protein is lethal in zebrafish and mice due to accelerated chromosome missegregation-induced apoptosis. Thus, the homozygous likely deleterious missense NM_022909.3:c.635T> C;p.(Ile212Thr) variant which we identified in one individual who presented with eight biochemical pregnancy losses and one fetal loss may be relevant as a germline-restricted phenotype akin to that observed with BUB1B (Baker et al. 2004). Interestingly, embryonic arrest was noted during her IVF treatment. Nineteen oocytes were obtained 12 of which were at metaphase II stage. Following injection, 11 oocytes were fertilized. Six embryos showed extensive fragmentation and arrest at different cleavage stage. Only one embryo reached the eight-cell stage and was transferred along with another seven-cell embryo, but no pregnancy was established.

    (v.) *PLK1*: Similarly to *CENPH, PLK1* is recruited to the kinetochore and is required for proper chromosomal segregation during cell division (Hood et al. 2012). Indeed, several papers describe an essential role by *PLK1* in oocyte development (Pahlavan et al. 2000; Solc et al. 2015; Tong et al. 2002). Again, complete deficiency is lethal and accompanied by increased aneuploidies (Lu et al. 2008), so we propose a germline-restricted phenotype in the individual with a homozygous variant NM_005030.3:c.128T>C;p.(Val43Ala). This individual presented with recurrent pregnancy loss. She had six intra-uterine inseminations, and while three were cancelled due to poor response, bleeding, and risk of hyper stimulation, she did conceive from one of the other three attempts. However, no fetal heartbeat was detected and the pregnancy was spontaneously aborted.

    (vi.) *PIF1*: The encoded helicase is highly conserved and is required to suppress DNA damage at G-quadruplex motifs (Paeschke et al. 2011). A very recent paper describes a *pif1*$^{-/-}$ fly model in which females failed to produce viable progeny (Kocak et al. 2019). This could be relevant to our finding of a homozygous NM_025049.3:c.506C>T;p.(Thr169Ile) variant in an individual with recurrent pregnancy loss (16 embryonic losses and 3 biochemical pregnancy losses).

    (vii.) *REXO4*: Very little is known about the encoded RNA exonuclease, but we opted to highlight the homozygous truncating variant NM_001279350.1:c.253C>T;p.(Gln85*) because of its role in estrogen receptor signaling (Krishnamurthy et al. 2011), which suggests a potential role in the affected individual with recurrent pregnancy loss (two biochemical pregnancy and one embryonic loss).

Discussion

The growing number of genes implicated in Mendelian forms of female infertility mirrors what has long been appreciated in animal models, i.e., many genes are required for normal fertility in females. While previous studies have focused on the role of ES in uncovering single-novel causes, this study seeks to provide a more comprehensive overview of the outcome of applying WES in women with various infertility phenotypes. We had excluded chromosomal aberrations, because these are routinely screened for, at least in women with recurrent miscarriage where they are seen in 12% of the couples, although not always uncovered by routine methods (Dong et al. 2019). Despite the relatively small cohort size, our approach revealed potential single-gene causes in a substantial minority (18.7%) not only in established genes, but also in genes with only tentative or even no link to human female reproduction. These variants spanned the spectrum of reproductive phenotypes from failure to produce oocytes to those with recurrent fetal losses, which increases the relevance of our findings to future studies of female infertility. Importantly, our results highlight that the distinction between the two groups of PI and RPL can be challenged. For example, when we first reported the identification of a founder Saudi variant in *TLE6*, the phenotype of the two original families was very consistent and comprised failure of the zygote to undergo first cleavage, thus representing the earliest known form of embryonic lethality in humans (Alazami et al. 2015). *TLE6* is one of several proteins that constitute the subcortical maternal complex (SCMC), and mutations in other components of SCMC were found to cause cleavage or early embryonic arrest following IVF (Wang et al. 2018). The woman which we report in this study with recurrent pregnancy loss and a homozygous truncating variant in *TLE6* demonstrates that oocytes depleted for *TLE6* have the capacity to undergo several post-fertilization divisions prior to arrest, which is consistent with what was observed in *Tle6*-/- mice (Yu et al. 2014).

The variant which we identified in *FSHR* in three sisters with failure of follicular growth and maturation is another example of phenotypic expansion. Thus far, only three OMIM phenotypes are listed for *FSHR*: ovarian dysgenesis 1 (OMIM: 233300), ovarian hyperstimulation syndrome (OMIM: 608115), and ovarian response to FSH stimulation (OMIM: 276400). Here, we observe a phenotype of normal looking ovaries that lack normal folliculogenesis despite stimulation with resulting loss of negative feedback and elevated LH (Luteinizing hormone) and FSH (follicle-stimulating hormone), but with normal secondary sexual characteristics. This overlaps with the phenotype originally reported in 1998 by Beau et al. who described an Armenian woman with secondary amenorrhea, high FSH levels, and normal-sized ovaries that lacked maturing follicles or corpus luteum, and was found to have a compound heterozygous mutation in *FSHR* (Beau et al. 1998). A third example of phenotypic expansion is our patient with a homozygous variant in *ZP1*. The typical phenotype of *ZP1*-related infertility is lack of zona pellucida (Huang et al. 2014). However, in our patient, no oocyte was retrieved after prolonged stimulation, thus confirming the empty follicle phenotype in described very recently (Dai et al. 2019; Sun et al. 2019; Yuan et al. 2019). Other examples of previously published links that we lend support to through our study are the novel homozygous variant in *NLRP5*, which was only recently linked to human female infertility with early embryonic arrest (Mu et al. 2019). Our patient's phenotype of embryonic arrest at the four-cell stage, therefore, seems to confirm this association. NLRP family proteins are abundant in oocytes, play critical roles in human reproduction, and are associated with multilocus imprinting disturbance. Biallelic mutation in *NLRP7* is known to cause recurrent hydatidiform mole (OMIM: 231090), and we were able to observe the formation of three pronuclei in oocytes from a woman with a homozygous *NLRP7* variant in this study.

Motile cilia have an established role in the clearing of respiratory epithelium and in sperm motility, the two main physiological functions that are lost in patients with primary ciliary dyskinesia (PCD). Despite the identical appearance of motile multicilia in the ovarian ducts to the respiratory epithelium, establishing female infertility as a phenotypic aspect of PCD has been challenging (Raidt et al. 2015). A large epidemiological study, however, lends credence to the original observations that female infertility is, indeed, more common in PCD patients (Vanaken et al. 2017). In fact, the two PCD genes we found mutated in women with infertility in this study (*DNAH11* and *CCNO*) were also found to be mutated in infertile women in that study. Our finding also seems to suggest that caution must be exercised before excluding the possibility of PCD in an infertile woman who lacks the typical respiratory presentation, which is also supported by the finding of PCD gene mutations in males with non-syndromic infertility.

In conclusion, our data support the use of WES to explore monogenic causes of female infertility beyond chromosomal causes. While this study was performed on a research basis, we believe that the growing number of monogenic causes of female infertility identified by this and other studies will eventually spur the clinical adoption of WES for these patients. The clinical utility of making a genetic etiology can be seen in the examples of monogenic forms that not amenable to treatment with IVF techniques, thus saving these women and their physicians' expensive IVF cycles that can be deemed futile by the genetic diagnosis. We also believe this line of research which will reveal novel drug targets for the development of new fertility regulators beyond the currently limited hormonal targets with their known side effects.

Acknowledgements We thank the study subjects for their enthusiastic participation. We thank the Genotyping and Sequencing Core Facilities at KFSHRC for their technical support.

Compliance with ethical standards

Conflict of interest Authors declare no conflict of interest.

References

Abe K, Naruse C, Kato T, Nishiuchi T, Saitou M, Asano M (2011) Loss of heterochromatin protein 1 gamma reduces the number of primordial germ cells via impaired cell cycle progression in mice. Biol Reprod 85:1013–1024

Alazami AM, Awad SM, Coskun S, Al-Hassan S, Hijazi H, Abdulwahab FM, Poizat C, Alkuraya FS (2015) TLE6 mutation causes the earliest known human embryonic lethality. Genome Biol 16:240

Alkuraya FS (2014) Genetics and genomic medicine in Saudi Arabia. Mol Genet Genom Med 2:369–378

Baker DJ, Jeganathan KB, Cameron JD, Thompson M, Juneja S, Kopecka A, Kumar R, Jenkins RB, De Groen PC, Roche P (2004) BubR1 insufficiency causes early onset of aging-associated phenotypes and infertility in mice. Nat Genet 36:744

Beau I, Touraine P, Meduri G, Gougeon A, Desroches A, Matuchansky C, Milgrom E, Kuttenn F, Misrahi M (1998) A novel phenotype related to partial loss of function mutations of the follicle stimulating hormone receptor. J Clin Investig 102:1352–1359

Beke A (2019) Genetic causes of female infertility. Genetics of endocrine diseases and syndromes. Springer, Berlin, pp 367–383

Brown JP, Bullwinkel J, Baron-Lühr B, Billur M, Schneider P, Winking H, Singh PB (2010) HP1γ function is required for male germ cell survival and spermatogenesis. Epigenet Chromat 3:9

Dai C, Chen Y, Hu L, Du J, Gong F, Dai J, Zhang S, Wang M, Chen J, Guo J, Zheng W, Lu C, Wu Y, Lu G, Lin G (2019) ZP1 mutations are associated with empty follicle syndrome: evidence for the existence of an intact oocyte and a zona pellucida in follicles up to the early antral stage. A case report. Hum Reprod 34:2201–2207

Dong Z, Yan J, Xu F, Yuan J, Jiang H, Wang H, Chen H, Zhang L, Ye L, Xu J (2019) Genome sequencing explores complexity of chromosomal abnormalities in recurrent miscarriage. Am J Hum Genet 20:20

Greenhouse S, Rankin T, Dean J (1998) Genetic causes of female infertility: targeted mutagenesis in mice. Am J Hum Genet 62:1282–1287

Group ECW (2008) Genetic aspects of female reproduction. Human Reprod Update 14:293–307

Guzeloglu-Kayisli O, Lalioti MD, Aydiner F, Sasson I, Ilbay O, Sakkas D, Lowther KM, Mehlmann LM, Seli E (2012) Embryonic poly (A)-binding protein (EPAB) is required for oocyte maturation and female fertility in mice. Biochem J 446:47–58

Hood EA, Kettenbach AN, Gerber SA, Compton DA (2012) Plk1 regulates the kinesin-13 protein Kif2b to promote faithful chromosome segregation. Mol Biol Cell 23:2264–2274

Huang H-L, Lv C, Zhao Y-C, Li W, He X-M, Li P, Sha A-G, Tian X, Papasian CJ, Deng H-W (2014) Mutant ZP1 in familial infertility. N Engl J Med 370:1220–1226

Huang N, Xia Y, Zhang D, Wang S, Bao Y, He R, Teng J, Chen J (2017) Hierarchical assembly of centriole subdistal appendages via centrosome binding proteins CCDC120 and CCDC68. Nat Commun 8:15057

Kocak E, Dykstra S, Nemeth A, Coughlin CG, Rodgers K, McVey M (2019) The *Drosophila melanogaster* PIF1 helicase promotes survival during replication stress and processive DNA synthesis during double-strand gap repair. Genetics 213:835–847

Krishnamurthy N, Ngam CR, Berdis AJ, Montano MM (2011) The exonuclease activity of hPMC2 is required for transcriptional regulation of the QR gene and repair of estrogen-induced abasic sites. Oncogene 30:4731

Lu L-Y, Wood JL, Minter-Dykhouse K, Ye L, Saunders TL, Yu X, Chen J (2008) Polo-like kinase 1 is essential for early embryonic development and tumor suppression. Mol Cell Biol 28:6870–6876

Maddirevula S, Coskun S, Alhassan S, Elnour A, Alsaif HS, Ibrahim N, Abdulwahab F, Arold ST, Alkuraya FS (2017a) Female infertility caused by mutations in the oocyte-specific translational repressor PATL2. Am J Hum Genet 101:603–608

Maddirevula S, Coskun S, Awartani K, Alsaif H, Abdulwahab F, Alkuraya F (2017b) The human knockout phenotype of PADI6 is female sterility caused by cleavage failure of their fertilized eggs. Clin Genet 91:344–345

Monies D, Abouelhoda M, AlSayed M, Alhassnan Z, Alotaibi M, Kayyali H, Al-Owain M, Shah A, Rahbeeni Z, Al-Muhaizea MA (2017) The landscape of genetic diseases in Saudi Arabia based on the first 1000 diagnostic panels and exomes. Human Genet 136:921–939

Mu J, Wang W, Chen B, Wu L, Li B, Mao X, Zhang Z, Fu J, Kuang Y, Sun X (2019) Mutations in NLRP2 and NLRP5 cause female infertility characterised by early embryonic arrest. J Med Genet 56:471–480

Nishimura K, Dioguardi E, Nishio S, Villa A, Han L, Matsuda T, Jovine L (2019) Molecular basis of egg coat cross-linking sheds light on ZP1-associated female infertility. bioRxiv 20:591966

Paeschke K, Capra JA, Zakian VA (2011) DNA replication through G-quadruplex motifs is promoted by the *Saccharomyces cerevisiae* Pif1 DNA helicase. Cell 145:678–691

Pahlavan G, Polanski Z, Kalab P, Golsteyn R, Nigg EA, Maro B (2000) Characterization of polo-like kinase 1 during meiotic maturation of the mouse oocyte. Dev Biol 220:392–400

Raidt J, Werner C, Menchen T, Dougherty GW, Olbrich H, Loges NT, Schmitz R, Pennekamp P, Omran H (2015) Ciliary function and motor protein composition of human fallopian tubes. Human Reprod 30:2871–2880

Solc P, Kitajima TS, Yoshida S, Brzakova A, Kaido M, Baran V, Mayer A, Samalova P, Motlik J, Ellenberg J (2015) Multiple requirements of PLK1 during mouse oocyte maturation. PLoS ONE 10:e0116783

Stephen EH, Chandra A (1998) Updated projections of infertility in the United States: 1995–2025. Fertil Steril 70:30–34

Sun L, Fang X, Chen Z, Zhang H, Zhang Z, Zhou P, Xue T, Peng X, Zhu Q, Yin M, Liu C, Deng Y, Hu H, Li N (2019) Compound heterozygous ZP1 mutations cause empty follicle syndrome in infertile sisters. Hum Mutat 40:2001–2006

Tong C, Fan H-Y, Lian L, Li S-W, Chen D-Y, Schatten H, Sun Q-Y (2002) Polo-like kinase-1 is a pivotal regulator of microtubule assembly during mouse oocyte meiotic maturation, fertilization, and early embryonic mitosis. Biol Reprod 67:546–554

Vanaken GJ, Bassinet L, Boon M, Mani R, Honore I, Papon J-F, Cuppens H, Jaspers M, Lorent N, Coste A (2017) Infertility in an adult cohort with primary ciliary dyskinesia: phenotype–gene association. Eur Respir J 50:1700314

Wang X, Song D, Mykytenko D, Kuang Y, Lv Q, Li B, Chen B, Mao X, Xu Y, Zukin V (2018) Novel mutations in genes encoding subcortical maternal complex proteins may cause human embryonic developmental arrest. Reprod Biomed Online 36:698–704

World Health Organization (2010) WHO laboratory manual for the examination and processing of human semen, 5th edn. World Health Organization, Geneva

Yu X-J, Yi Z, Gao Z, Qin D, Zhai Y, Chen X, Ou-Yang Y, Wang Z-B, Zheng P, Zhu M-S (2014) The subcortical maternal complex controls symmetric division of mouse zygotes by regulating F-actin dynamics. Nat Commun 5:4887

Yuan P, Li R, Li D, Zheng L, Ou S, Zhao H, Zhang Q, Wang W (2019) Novel mutation in the ZP1 gene and clinical implications. J Assist Reprod Genet 36:741–747

Zegers-Hochschild F, Adamson GD, Dyer S, Racowsky C, de Mouzon J, Sokol R, Rienzi L, Sunde A, Schmidt L, Cooke ID, Simpson JL, van der Poel S (2017) The international glossary on infertility and fertility care, 2017. Hum Reprod 32:1786–1801

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

De La Vega *et al. Genome Medicine* (2021) 13:153
https://doi.org/10.1186/s13073-021-00965-0

Genome Medicine

RESEARCH

Open Access

Artificial intelligence enables comprehensive genome interpretation and nomination of candidate diagnoses for rare genetic diseases

Francisco M. De La Vega[1,2,3], Shimul Chowdhury[4], Barry Moore[5], Erwin Frise[1], Jeanette McCarthy[1], Edgar Javier Hernandez[5], Terence Wong[4], Kiely James[4], Lucia Guidugli[4], Pankaj B. Agrawal[6,7], Casie A. Genetti[6], Catherine A. Brownstein[6], Alan H. Beggs[6], Britt-Sabina Löscher[8], Andre Franke[8], Braden Boone[9], Shawn E. Levy[9], Katrin Õunap[10,11], Sander Pajusalu[10,11], Matt Huentelman[12], Keri Ramsey[12], Marcus Naymik[12], Vinodh Narayanan[12], Narayanan Veeraraghavan[4], Paul Billings[1], Martin G. Reese[1*], Mark Yandell[1,5*] and Stephen F. Kingsmore[4]

Abstract

Background: Clinical interpretation of genetic variants in the context of the patient's phenotype is becoming the largest component of cost and time expenditure for genome-based diagnosis of rare genetic diseases. Artificial intelligence (AI) holds promise to greatly simplify and speed genome interpretation by integrating predictive methods with the growing knowledge of genetic disease. Here we assess the diagnostic performance of Fabric GEM, a new, AI-based, clinical decision support tool for expediting genome interpretation.

Methods: We benchmarked GEM in a retrospective cohort of 119 probands, mostly NICU infants, diagnosed with rare genetic diseases, who received whole-genome or whole-exome sequencing (WGS, WES). We replicated our analyses in a separate cohort of 60 cases collected from five academic medical centers. For comparison, we also analyzed these cases with current state-of-the-art variant prioritization tools. Included in the comparisons were trio, duo, and singleton cases. Variants underpinning diagnoses spanned diverse modes of inheritance and types, including structural variants (SVs). Patient phenotypes were extracted from clinical notes by two means: manually and using an automated clinical natural language processing (CNLP) tool. Finally, 14 previously unsolved cases were reanalyzed.

---

* Correspondence: mreese@fabricgenomics.com; myandell@genetics.utah.edu
[1]Fabric Genomics Inc., Oakland, CA, USA
Full list of author information is available at the end of the article

© The Author(s). 2021 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Results: GEM ranked over 90% of the causal genes among the top or second candidate and prioritized for review a median of 3 candidate genes per case, using either manually curated or CNLP-derived phenotype descriptions. Ranking of trios and duos was unchanged when analyzed as singletons. In 17 of 20 cases with diagnostic SVs, GEM identified the causal SVs as the top candidate and in 19/20 within the top five, irrespective of whether SV calls were provided or inferred ab initio by GEM using its own internal SV detection algorithm. GEM showed similar performance in absence of parental genotypes. Analysis of 14 previously unsolved cases resulted in a novel finding for one case, candidates ultimately not advanced upon manual review for 3 cases, and no new findings for 10 cases.

Conclusions: GEM enabled diagnostic interpretation inclusive of all variant types through automated nomination of a very short list of candidate genes and disorders for final review and reporting. In combination with deep phenotyping by CNLP, GEM enables substantial automation of genetic disease diagnosis, potentially decreasing cost and expediting case review.

Background

A central tenet of genomic medicine is that outcomes are improved when symptom-based diagnoses and treatments are augmented with genetic diagnoses and genotype-differentiated treatments. Worldwide, an estimated 7 million infants are born with serious genetic disorders every year [1]. The last decade witnessed a huge increase in the catalog of genes associated with Mendelian conditions, from about 2300 in 2010 [2], to over 6700 by the end of 2020 [3]. The translation of that knowledge, in conjunction with major improvements in WES and WGS and downstream analytical pipelines, has enabled increased rates of diagnosis, from about 10%, with single gene tests, to over 50% [4]. While limitations of read alignment and variant calling were major obstacles to early clinical implementations of WES and WGS [5], they have been largely removed by algorithmic advances, hardware acceleration, and parallelization through cloud computing [6, 7]. However, clinical interpretation of genetic variants in the context of the patient's phenotype remains largely manual and extremely labor-intensive, requiring highly trained expert input. This remains a major barrier to widespread adoption and contributes to continued low rates of genomic testing for patients with suspected genetic disorders despite strong evidence for diagnostic and clinical utility and cost effectiveness [8].

The major challenge for genome-based diagnosis of rare genetic disease is to identify a putative disease-causing variant amid approximately four million benign variants in each genome, a problem akin to finding a needle in a haystack [9]. Clinical genome interpretation is, by necessity, performed by highly trained, scarce, genome analysts, genetic counselors, and laboratory directors [10]. For an average of 100 variants for review per case [11], this translates to 50–100 h of expert review per patient [10]. In practice, this has led to review of only about 10 variants per case, which somewhat defeats the purpose of genome-wide sequencing.

The genome interpretation process consists of iterative variant filtering, coupled with evidence-based review of candidate disease-causing variants [12]. This process was almost entirely manual until the advent of variant prioritization algorithms, such as Annovar [13] and VAAST [14], and was later improved by the integration of patient phenotypes in analyses, e.g., Phevor [15], Exomiser [16], Phen-Gen [17], Phenolyzer [18], and more recently Amelie [19]. While these tools accelerate review times, their stand-alone performance has been insufficient for widespread clinical adoption, in part due to their inability to appropriately interpret structural variants (SVs). SVs account for over 10% of Mendelian disease [20, 21], and about 20% of diagnoses in routine neonatal intensive care unit (NICU) [22] and pediatric patients [23]. Unified methods for prioritization of SVs, SNVs, and small indels are a fundamental requirement for further automation of genome interpretation.

The use of artificial intelligence (AI) has made significant inroads in healthcare [24], and a new class of genome interpretation methods [19, 25–28] are being developed with the promise of removing the interpretation bottleneck for rare genetic disease diagnosis through electronic clinical decision support systems (eCDSS) [29]. Speed and accuracy of interpretation are particularly important for seriously ill children in the NICU [27], where diagnosis in the first 24–48 h of life has been shown to maximally improve health outcomes [30]. The settings and extent to which AI facilitates diagnosis are still being investigated [27, 28]. Issues include what types of AI methods are most suitable (e.g., Bayesian networks, decision trees, neural nets [31]); how they compare with current variant prioritization approaches in terms of accuracy; their diagnostic performance across different clinical scenarios and variant types; their potential to offer new forms of decision support; and how well they integrate with automated patient phenotyping and clinical decision making [27, 28, 32].

Algorithmic benchmarking in this domain is no simple matter. Hitherto, most attempts have used simulated cases (created by adding known disease-causing variants to reference exomes and genomes), included only a few cases, derived from a single center, or were limited to certain variant types [17, 33, 34]. Such benchmarking is inherently limited, as it is not representative of the true diversity of genetic diseases and variant types (e.g., by omitting cases with causal SVs), and provide no means to evaluate the impact of different sequencing and variant calling pipelines on performance.

Here we describe and benchmark the diagnostic performance of Fabric GEM (hereafter referred to as "GEM"), a new AI-based eCDSS, and compare it to current variant prioritization approaches using a diverse cohort of retrospective pediatric cases from the Rady Children's Institute for Genomic Medicine (RCIGM). These cases are largely comprised of seriously ill NICU infants; all were diagnosed with Mendelian conditions following WGS (or, in a few cases, WES), using a combination of filtering and variant prioritization approaches. These real-world cases encompass the breadth of phenotypes and disease-causing variants, including pathogenic SVs. We then sought to replicate the diagnostic performance of GEM in a second set of affected, diagnosed, and undiagnosed children outside the NICU. They were collected from five different academic medical centers, mostly consisting of WES, to examine the generalizability of GEM's diagnostic performance to other sequencing, variant calling pipelines, and clinical settings. Finally, we reanalyzed a set of previously negative RCIGM cases to evaluate the ability of GEM to identify new diagnoses without suggesting numerous false positives that would lead to time-consuming case reviews. Our results show that rapid, accurate, and comprehensive WGS- and WES-based diagnosis is achievable through integration of new data modalities with algorithmic innovations made possible by AI.

Methods

Patient selection, phenotyping, and specimen sequencing

This retrospective study was designed to provide benchmark data to test the GEM eCDSS. We compiled 119 cases from Rady Children's Hospital (the *Benchmark* cohort), consisting of mostly NICU admissions, and 60 additional cases from five academic medical centers (the *Validation* cohort), which consisted mostly of referrals from genetic clinics and none included causal structural variants, as described below.

Rady Children's Hospital

In total, 119 cases with primary findings, deemed definitively solved using previously published methods [27, 30, 35], and 14 negative cases, were sequenced as part of the rapid-WGS (rWGS) sequencing program at the Rady Children's Hospital Clinical Genome Center. These cases where a sample of convenience, drawn from the first symptomatic children who were enrolled in four previously published studies that examined the diagnostic rate, time to diagnosis, clinical utility of diagnosis, outcomes, and health care utilization of rWGS between 26 July 2016 and 25 September 2018 at Rady Children's Hospital, San Diego, USA. One of the studies was a randomized controlled trial of genome and exome sequencing (ClinicalTrials.gov identifier: NCT03211039) [30]; the others were cohort studies (ClinicalTrials.gov identifiers: NCT02917460, and NCT03385876) [35–40]. All subjects had a symptomatic illness of unknown etiology in which a genetic disorder was suspected, had a Rady Children's Hospital Epic EHR, and that had clinical phenotype descriptions expressed as human phenotype ontology terms both manually curated by clinicians and automatically extracted by CNLP (Additional file 1: Table S1).

WGS (or in a few instances WES) was performed as previously described [35, 40]. Briefly, PCR-free WGS was performed to an average of 40× coverage in the Illumina HiSeq 2000, HiSeq 4000, and NovaSeq 6000 sequencers. Alignment and sequence variant calling were performed using the Illumina DRAGEN software, while copy number variation was identified through an approach that integrates the tools Manta [41] and CNVnator [42]. Structural variants were then filtered for recurrent artifacts observed in previous non-affected cases and only included in the input VCF file if they overlap a known disease gene (OMIM). All variants reported as primary findings were validated orthogonally by Sanger sequencing. In the case of trios, de novo origin of reported variants was established by comparing to their parents' data. In some older cases, SV calling was not performed; any causal SVs therein were identified by an orthogonal CGH microarray or manual inspection of alignments. In what follows, we refer to these 119 cases with primary findings as the *Benchmark* cohort, and the 14 negative cases as the *Unsolved* cohort.

Boston Children's Hospital

Eleven cases (all single probands) from the Beggs Lab, Congenital Myopathy Research Program laboratory, and Manton Center for Orphan Disease Research at Boston Children's Hospital were included in the analysis [43–48].

Libraries (TruSeq DNA v2 Sample Preparation kit; Illumina, San Diego, CA) and whole-exome capture (EZ Exome 2.0, Roche) were performed according to manufacturer protocols from DNA extracted from blood samples. WES was carried out on an Illumina HiSeq 2000. Reads were aligned to the GRCh37/hg19 human genome assembly using an in-house assembler. Variants were called using Gene Analysis Toolkit (GATK) version 3.1 or higher (Broad Institute, Cambridge, MA) and were Sanger confirmed by the Boston Children's Hospital IDDRC Molecular Genetics Core Facility.

Christian-Albrechts University of Kiel

Twelve cases (all single probands) from the Institute of Clinical Molecular Biology (IKMB) were included in the analysis [49–55].

Illumina's Nextera/TruSeq whole-exome target capture method was applied. WES was carried out on the Illumina HiSeq/NovaSeq platforms. Reads were aligned to the GRCh37/hg19 human genome assembly using BWA-MEM version 0.7.17 and variants called using GATK version 4.1.6.0 (Broad Institute, Cambridge, MA).

HudsonAlpha Institute for Biotechnology

Three cases (two trios and a single proband) from the Clinical Services Laboratory at HudsonAlpha Institute for Biotechnology, including cases from the Clinical Sequencing Evidence-Generating Research (CSER) consortium, were included in the analysis [56–59].

WGS was carried out on an Illumina HiSeq X. Reads were aligned to the GRCh37/hg19 human genome assembly followed by variant calling using the Illumina DRAGEN software version 3.2.8 (Illumina, Inc. San Diego, CA).

Translational Genomics Research Institute

Twenty-three cases (including singletons, duos, trios, and quads) from the Center for Rare Childhood Disorders at The Translational Genomics Research Institute (TGen) were included in the analysis [60, 61].

WES or WGS sequencing was carried out on an Illumina HiSeq 2000, HiSeq 2500, HiSeq 4000, or NovaSeq6000. For WES, the Agilent SureSelect Human All Exon V6 or CRE V2 target capture method was applied. Reads were aligned to reference GRCh37 version hs37d5 and variants called using GATK Haplotype caller version 3.3-0-g37228af (Broad Institute, Cambridge, MA).

Tartu University Hospital

Eleven cases from Tartu University Hospital in Estonia that had undergone WES were included in the analysis [62–64].

Nextera Rapid Capture Exome Kit-i (Illumina Inc.) target capture method was applied. WES was carried out on an Illumina HiSeq2500 sequencer. Reads were aligned to the GRCh37/hg19 human genome assembly using BWA-MEM version 0.5.9 and variants called using GATK Haplotype caller version 3.4 (Broad Institute, Cambridge, MA).

Variant annotation and data sources

All analyses were performed based on the GRCh37 human genome assembly. Variant consequences and annotations were obtained with VEP v.95 [65] utilizing ENSEMBL transcripts version 95 (excluding non-coding transcripts) and selecting the canonical transcript for analysis. Transcript-specific prediction for evaluating variant deleteriousness was calculated with VVP [66], which were also used as input for VAAST [14]. Variants were annotated with ClinVar (version 20200419) [67] ensuring exact position and base match. Gene conditions were extracted from OMIM (version 2020_07) [68] and HPO (obo file dated 2020-08-11) [69]. Gene symbols were harmonized using ENSEMBL and HGNC databases controlling for synonymous gene symbols.

AI-based disease gene and condition prioritization

AI-based prioritization and scoring of candidate disease genes and diagnostic conditions was performed using Fabric GEM [70], which is a commercially available eCDSS part of the Fabric Enterprise platform (Fabric Genomics, Oakland, CA). GEM inputs are genetic variant calls in VCF format and case metadata, including (optional) parental affection status, and patient (proband) phenotypes in the form of Human Phenotype Ontology (HPO) terms. The VCF files can include "small variants" (single nucleotide, multiple nucleotide, and small insertion/deletion variants), and (optionally) structural variants (insertion/deletions of over 50 bp, inversions, and/or copy number variants with imprecise ends). This information can be provided via an application programming interface or manually in the user interface. Data analysis is typically carried out in minutes depending on inputs. GEM outputs are displayed in an interactive report (Additional file 2: Figure S1) that includes a list of candidate genes ranked by the GEM gene score (see below), detailed information of patient variants present in each candidate gene, and conditions associated with each candidate gene ranked by GEM's condition match (CM) score (explained below).

GEM aggregates inputs from multiple variant prioritization algorithms with genomic and clinical database annotations, using Bayesian means to score and prioritize potentially damaged genes and candidate diseases. Briefly, the algorithm parametrizes itself using the proband's called variants as one-time, run-time training data, inferring the states of multiple variables directly from the input variant distribution, e.g., sex. Additional static training parameters were derived from the 1000 Genomes Project [71] and CEPH [72] genome datasets. GEM reevaluates genotype calls and quality scores considering read support, genomic location, proband sex, and potentially overlapping SVs, augmenting the genotype calls with more nuanced posterior probabilities, computing ploidy for each variant. GEM also computes the likelihood that the proband belongs to any of several different ancestry groups using the input genotypes together with gnomAD sub-population variant frequency data [73]. The probabilities of other, internal, variables, conditioned on each state (sex and ancestry, etc.) are then obtained using naive Bayes, controlling for non-independence of variables by calculating a correlation matrix at run time using the proband's data. For example, after conditioning variant scores on ancestry, known inheritance pattern for the gene in question, gene location, and proband sex, GEM may conclude that a de novo variant is unlikely to participate in a disease-causing genotype, even though it is predicted to be highly damaging. Thus, highly damaging and de novo variants, even frameshifting ones, do not automatically receive high GEM scores. GEM uses the same procedure to evaluate and score biallelic genotypes for known and novel disease-gene candidates. The only difference is that the global prior (e.g., relative proportion of known disease genes with autosomal recessive vs. autosomal dominant inheritance patterns), rather than OMIM and HPO support for a particular inheritance pattern at that locus, is used to evaluate possible biallelic cases in novel gene candidates.

GEM's gene scores are Bayes factors (BF) [74]. Analogous to the likelihood ratio test, a Bayes factor presents the $\log_{10}$ ratio between the posterior probabilities of two models, summarizing the relative support for the hypothesis (in this case) that the prioritized genotype damages the gene in which it resides and that explains the proband's phenotype versus the contrapositional hypothesis that the variant neither damages the gene nor explains the proband's phenotype. In keeping with established best practice [74], a $\log_{10}$ Bayes factor between 0 and 0.69 is considered moderate support, between 0.69 and 1.0 substantial support, between 1.0 and 2.0 strong support, and above 2.0, decisive support. A score less than 0 indicates that the counter hypothesis is more likely. For calculating the Bayes posterior $p(M|D)$, the probability of the data given a model $(pD|M)$ is derived from GEM's severity scoring protocol, which includes input from the VAAST and VVP algorithms, and any available prior variant classifications from the ClinVar database. This model is conditioned upon sex, ancestry, feasible inheritance model, gene location, and gene-phenotype priors derived by seeding the provided patient HPO terms to the HPO ontology graph and subsequently obtaining priors for all genes in the HPO and GO ontologies by belief propagation using Phevor's previously described Bayesian network-based algorithm [15]. The prior probability for the model (pM) is based upon known disease associations in the Mendelian conditions databases OMIM and/or HPO with the gene in question.

GEM's Bayes factor-based scoring system is designed for ease of explanation and to speed interpretation. GEM scores are not intended to be definitive, rather they are designed to provide guidance for succinct case reviews carried out by clinical geneticists. Thus, GEM outputs also include several additional scores that provide additional guidance and improve explainability. GEM gene scores, for example, are accompanied by VAAST [14], VVP [66], and Phevor [15] posterior probabilities, conditioned upon the potentially confounding variables of proband sex, gene location, and ancestry, together with common variant genomic and clinical annotations (Additional file 2: Figure S1). These scores further ease interpretation, as they allow users to assess the major drivers of a GEM score and their relative contributions to it.

GEM also provides means to assess the Mendelian conditions associated with putative disease-causing genes as possible diagnoses via its condition match (CM) scores. Like gene scores, CM scores are Bayes factors and are derived from the $\log_{10}$ ratio of the posterior probability that HPO phenotype associations for a given Mendelian condition's HPO are consistent with the proband's phenotype versus the contrapositional hypothesis. For these calculations, the probability of the data, $p(D|M)$, is determined using Phevor's Bayesian algorithm to obtain a probability for each disease, conditioned upon the proband's phenotype. The prior probability for the model, $p(M)$, is the probability that one or more genes associated with the Mendelian condition (as documented in OMIM and/or HPO) contain a damaging genotype as ascertained by GEM's severity scoring protocol. Condition match scores are displayed alongside each gene-associated condition for review (Additional file 2: Figure S7).

Structural variant scoring and ab initio inference by GEM

At run time, GEM infers ab initio the existence of SVs, their coordinates, and their copy numbers (ploidy) in a probabilistic fashion using SNVs, sort indel calls, read depths, zygosity, and gnomAD frequency data. GEM searches the proband's genotypes for evidence of three types of SV: deletions, duplications, and CNVs. Regions exhibiting loss of heterozygosity (LOH), for example, are used as evidence for heterozygous deletions. Genomic spans lacking expected variants, the signature of homozygous deletions, are identified using gnomAD population frequencies [73] to derive point estates that a given gnomAD variant would or would not be ascertained given its population frequency. Further evidence for duplications and deletions is derived from read support, e.g., approximately integer increases or decreases in depth across a region provide support for copy number variation. Point estimates at each site of a small variant call are further conditioned upon provided variables, such as genotype qualities, and inferred ones, such as sex and ancestry, to obtain more refined estimates. High scoring segments and their maximum likelihood start and end coordinates are identified using a Markov model [75]. The results are used to determine the degree of support for external SV calls, and as the basis for GEM's own SV calls. For ease of reporting, ab initio SV calls that overlap an external SV call (default minimum reciprocal overlap of 33%) are replaced in the output by the external SV call as long as they still overlap the relevant scored genes.

Benchmarking variant prioritization with VAAST, Phevor, and Exomiser

We used the Snakemake software [76] to create a workflow that analyzes cases with the VAAST, Phevor, and Exomiser algorithms. This workflow was only applied to the benchmark cohort to enable us to compare the performance of four genome interpretation tools with exactly the same inputs and annotations. The pipeline starts with a VCF file, family structure, affection status, and HPO terms and concludes with the outputs for each of the algorithms. VVP scores were obtained as described above and provided to VAAST as input. VAAST was provided pedigree information and affection status and was run in both dominant and recessive modes with results aggregated. Gene ranks for VAAST are reported for the highest scoring occurrence of the gene from aggregated outputs. Phevor was provided with HPO terms and VAAST scores as inputs. Ranks were selected as described for VAAST.

Exomiser [16] benchmark analyses were run with the same configuration used in the 100,000 Genomes Project [77], specifically (1) using the GRCh37 genome assembly; (2) analyzing autosomal and X-linked forms of dominant and recessive inheritance; (3) allele frequency sources from the 1000 Genomes Project [78], TopMed [79], UK10K [80], ESP, ExAC [81], and gnomAD [73] (except Ashkenazi Jewish); (4) pathogenicity sources from REVEL [82] and MVP [83]; and (5) including the steps failedVariantFilter, variantEffectFilter (remove non-coding variants), frequencyFilter with maxFrequency = 2.0, pathogenicityFilter with keepNonPathogenic = true, inheritanceFilter, omimPrioritiser, and hiPhivePrioritiser.

Exomiser was considered to have identified the diagnosed gene when it was scored as a candidate for any of the utilized modes of inheritance. None of the tools in this analysis were provided a target mode of inheritance (as it is unknown), and so the diagnostic gene rank for Exomiser was determined from its rank within the combined gene candidate list from all modes of inheritance (i.e., the same procedure used for VAAST and PHEVOR). The ranks within the combined list of candidate genes were generated by sorting gene-level candidates from all modes of inheritance on the Exomiser combinedScore in descending order with each candidate gene only added to the list on its first, highest scoring occurrence. Exomiser variant level scoring was not considered for determining candidates or ranking. All Exomiser analyses on the benchmark cohort ran to completion and successfully produced output; however, in 18 cases, Exomiser did not identify the true positive diagnostic gene as a scored candidate (i.e., it was absent from its output). A similar phenomenon was observed in 4 cases using VAAST. For both tools, these cases were considered false negatives.

Impact of deep phenotypes derived from clinical NLP

The utility of HPO terms was investigated by rerunning all analyses from the benchmark cohort with three sets of HPO terms. The motivations for these analyses were first to determine how sensitive GEM is to phenotyping errors; and second, to compare the utility of CNLP-derived descriptions to manual ones. For each case, an HPO terms list was provided that included HPO terms manually curated by the analysis team when the case was originally solved. A second set of HPO terms was generated from NLP analysis of clinical notes related to the case using the CLiX ENRICH software (Clinithink, Alpharetta, GA) [28]. A randomized set of HPO terms was generated for each case whereby the number of HPO terms from the CliniThink analysis case was held constant, and alternate terms were randomly selected from the entire corpus of HPO terms across all samples with each selection probability determined by the number of times that term occurred in the corpus.

Results and discussion

GEM AI outperforms variant prioritization approaches

We benchmarked GEM, an AI-based eCDSS, using a cohort of 119 pediatric retrospective cases from Rady Children's Institute for Genomic Medicine (RCIGM; benchmark cohort). Most of these were critically ill NICU infants who received genomic sequencing for diagnosis of genetic diseases. All had been diagnosed with one or more Mendelian conditions using a combination of manual filtering and variant prioritization approaches ("Methods"). To further validate performance, we also analyzed a second cohort comprised of 60 non-NICU, rare disease patients from five different academic medical centers (validation cohort). Finally, we reanalyzed a set of 14 previously analyzed probands that had remained undiagnosed by WGS. Our goal was to evaluate the ability of GEM to identify new diagnoses in these previously unsolved cases, without providing false positive findings that would result in time-consuming case reviews. To provide context for our performance benchmarks, we also ran three commonly used variant prioritization tools: VAAST [14], Phevor [15], and Exomiser [16].

The benchmark and validation cohorts included singleton probands, parent-offspring trios, different modes of inheritance, and both small causal variants (SNVs, and small insertions or deletions, indels; Table 1; Additional file 1: Table S1) and large structural variants (SV), some of which were causative (Table 2). In these retrospective analyses, we considered the variants, disease genes, and conditions that were included as primary findings in the clinical report as the "gold standard" truth set.

GEM gene scores are Bayes factors (BF) [84]; these were used to rank gene candidates (Additional file 2: Figure S1). BFs are widely used in AI, as they concisely quantify the degree of support for a conclusion derived from diverse lines of evidence. In keeping with established practices [84], a BF of 0–0.69 was considered moderate support, 0.69–1.0 substantial support, 1.0–2.0 strong support, and above 2.0, decisive support [84]. Scores less than 0 indicated support for the counter hypothesis—that variants in that gene were not causal for the proband's disease. GEM outputs also include several annotations and metrics that provide additional, supportive guidance for subsequent expert case review (Additional file 2: Figure S1). Experience has shown that such guidance is critical for adoption by experts who wish to review the evidence supporting automated variant assertions. These include VAAST, VVP, and Phevor posterior probabilities, conditioned upon proband sex, gene location, and ancestry. Annotations include variant consequence, ClinVar database pathogenicity assessments, and OMIM conditions associated with genes. This metadata enables expert users to review the major contributions underpinning a final GEM score. Moreover, GEM prioritizes diplotypes, rather than variants, which speeds interpretation of compound heterozygous variants in recessive diseases (Additional file 2: Figure S1B). Comparison of the diagnostic performance of GEM to variant prioritization methods utilized ranking of the correct diagnostic gene. We assumed that in the case of compound heterozygotes, variant prioritization methods such as Exomiser would rank one variant of the pair highly, leading to identification of the other upon manual review ("Methods").

GEM ranked 97% of previously reported causal gene(s) and variant(s) among the top 10 candidates in the 119 benchmark cohort cases. In 92% of cases, it ranked the correct gene and variant in the top 2 (Fig. 1A). By comparison, the next best algorithm, Phevor, identified 73% of causal variants in the top 10 candidates and 59% in the top 2. GEM, Phevor, and Exomiser prioritize results by patient phenotypes (provided as HPO terms) in addition to variant pathogenicity, whereas VAAST only utilizes genotype data, explaining its lower performance. Thus, these data also highlight that patient phenotypes improve the diagnostic performance of automated interpretation tools.

The benchmark cohort included 3 cases for which two genes were reported to contribute to the patient phenotype. This rate (2.5%) is consistent with previous reports for digenic inheritance [85]. The statistics above use the top ranked genes in these cases, but Additional file 1: Table S3 shows that GEM also ranked the second causal gene among its top candidates, whereas Phevor reported poor ranks in one case, and Exomiser missed the second gene in two out of the three cases.

Table 1 Characteristics of case cohorts. Benchmark cohort, 119 cases total. Validation cohort, 60 cases total. Grand total, 179 cases

	Assay type		Variant type		Proband sex		Pedigree Type		
Mode of inheritance	**WGS**	**WES**	**SNV/Indel**	**SV**	**Male**	**Female**	**Single**	**Duos**	**Trios**
Benchmark cohort									
Autosomal dominant	70	11	66	15	36	45	35	6	40
Autosomal recessive	27	–	23	4	14	13	9	1	17
X-linked dominant	6	–	5	1	1	5	2	–	4
X-linked recessive	5	–	5	–	5	–	2	1	2
Sub-totals	108	11	99	20	56	63	48	8	63
Validation cohort									
Autosomal dominant	3	34	37	–	10	27	15	2	20
Autosomal recessive	1	14	15	–	5	10	9	–	6
X-linked dominant	1	5	6	–	3	3	1	3	2
X-linked recessive	–	2	2	–	2	–	1	–	1
Sub-totals	5	55	60	0	20	40	26	5	29

Table 2 Diagnostic structural variants identified by GEM in the benchmarking cohort (20 out of 119 cases). Structural variants are ranked by GEM based on the genes harbored by the variant and presented alongside other ranked genes with coding SNVs or small indels based on the top scored gene. The asterisk indicates genes that in the literature are candidates for the phenotype of the diagnostic disease/syndrome (as described in OMIM). The results show that GEM can analyze both deletions (del) and duplications (dup) of sizes as small as 4 kb and up to entire chromosome arms, diverse modes of inheritance, pedigree structure, and from either WGS or WES assay data. GEM also automatically identified compound heterozygotes between SVs and SNV/indels (cases 1, 2, and 8). Input SV calls can include breakpoint-based calls (here "SV"), or imprecise CNV calls based on read depth analysis. Notably, GEM can also infer SVs directly from the small variant data when external SV calls are not provided (cases 2, 10, 15, and 17), and score them appropriately, identifying diagnostic variants that in the original cases were found by microarrays and not by sequencing

Case no.	Top scored gene(s)	Gene rank	GEM score	Variant(s) position	SV type	Length (kb)	Mode of Inheritance	Pedigree type	Assay type	SV calls in input	Diagnosis
252268	FANCA*	1	2.28	chr16:89847864-89863349; FANCA: c.3788_3790delTCT	Del	15	Recessive	Trio	WGS	SV	Fanconi anemia
223449	TANGO2*	1	2.13	chr22:20028937-20057143; TANGO2: c.605+1G>A	Del	28	Recessive	Trio	WGS	None	MECRCN
266523	BTRC*	1	2.05	chr10:102941001-103430600	Dup	490	Dominant	Duo	WGS	SV	Split hand/foot malformation type 3
267392	HIRA, TBX1*	1	2.05	chr22:18893883-21562619	Del	2669	Dominant	Single	WES	CNV	DiGeorge syndrome; velocardiofacial syndrome
267148	KMT2A	1	1.87	chr11:116691508-126432828; chr22:17038511-20307516	Dup	9741; 3269	Dominant	Trio	WES	CNV	Emanuel syndrome
253691	HIRA, TBX1*	1	1.73	chr22:18893883-20307516	Del	1414	Dominant	Single	WES	CNV	DiGeorge syndrome; velocardiofacial Syndrome
256943	MAGEL2*	1	1.64	chr15:22833478-28566610	Del	5733	Dominant	Single	WES	CNV	Prader Willi syndrome
254012	NDUFS3*	1	1.56	chr11:47605229-47609177; NDUFS3: c.374G>A	Del	4	Recessive	Trio	WGS	SV	Leigh syndrome
254728	EPHA4	2	1.46	chr2:220309089-224580863	Del	4272	Dominant	Single	WGS	SV	Pathogenic deletion in 2q35q36.1
44671	NPAP1	1	1.42	chr15 tetrasomy (broken in multiple dups)	Dup	4542; 991; 358; 158	Dominant	Trio	WGS	None	Isodicentric chromosome 15 syndrome
360547	FREM1	1	1.33	chr9:1-18477200	Del	18,437	Dominant	Trio	WGS	SV	Chromosome 9p deletion syndrome
259685	TYROBP	1	1.31	chr19:23158251-33502767	Dup	10,345	Dominant	Trio	WES	SV	Partial trisomy 19p12.q13.11
266700	TAB2	1	1.31	chr6:144951601-150260400	Del	5309	Dominant	Trio	WGS	SV	Chromsome 6q24-q25 Syndrome
244102	MAGEL2*	1	1.28	chr15:23684685-26108259	Del	2424	Dominant	Single	WES	CNV	Prader Willi syndrome
204560	JAG1*	2	1.21	chr20:10471400-13459333	Del	44	Dominant	Trio	WGS	None	Alagille syndrome
246146	HCN1	1	1.20	chr5:213101-46,270,700	Dup	44	Dominant	Single	WGS	SV	Trisomy 5p
45020	PCDH19*	1	1.15	chrX:92925011-99669272	Del	6744	X-linked dominant	Trio	WGS	None	Developmental and epileptic encephalopathy 9
248678	FANCC*	1	1.14	chr9:97998556-98009092	Del	11	Recessive	Single	WGS	SV	Fanconi Anemia
352726	THRA	1	1.00	chr17:32147833-79020944	Dup	46,873	Dominant	Proband	WGS	SV	Distal trisomy 17q
251355	TRIP11	4	0.58	chr14:84783523-96907490	Del	12,124	Dominant	Duo	WGS	SV	Chromosome 14q31.2q32.2 Syndrome

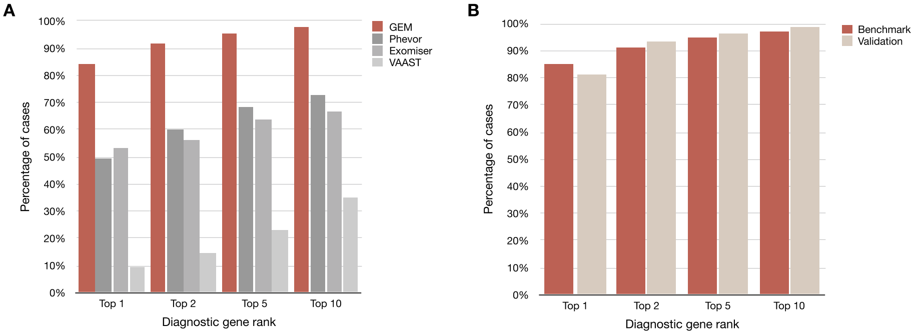


Fig. 1 The diagnostic sensitivity of GEM was greater than the variant prioritization methods Phevor, Exomiser, and VAAST. **A** Proportion of the benchmark cohort of 119 cases where the true causal genes (or variants in the case of causal SVs) were identified among the top 1st, 2nd, 5th, or 10th gene candidates. Patient phenotypes were extracted manually from medical records by clinicians and provided as HPO term inputs to GEM, Exomiser, and Phevor. VAAST only considers variant information. It should be noted that GEM and Phevor ranks correspond to genes, which may include one or two variants (the latter in the case of a compound heterozygote), whereas Exomiser and VAAST ranks were for single variants. In the case of compound heterozygotes, the rank of the top-ranking variant is shown for Exomiser and VAAST. **B** Comparison of GEM performance in the validation cohort (excluding SV cases) versus the validation cohort (comprised of 60 rare genetic disease cases from multiple sources)

Next, we investigated whether the diagnostic performance of GEM extended to Mendelian diseases other than those of NICU infants, such as patients with later disease onset, less severe presentations, or with data produced by other variant calling pipelines or outpatient genetic clinics. For these analyses, we compiled a validation cohort largely consisting of WES cases from five different academic medical centers (Table 1; Additional file 1: Table S2). The diagnostic performance of GEM in the validation cohort was almost identical to that in the benchmark cohort (Fig. 1B). These data demonstrated that the diagnostic performance of GEM was not dependent of disease severity, age of onset, or genomic sequencing or variant detection methods.

An implication of these findings is that GEM achieved 97% recall (true positive rate) by review of 10 genes, whereas the other tools had ≤ 78% recall by similar review (Fig. 1, Additional file 2: Figure S2). In part, this difference reflected the unique ability of GEM to prioritize SVs. Excluding SV cases, GEM, Phevor, and Exomiser achieved recall of 97%, 83%, and 76%, respectively, by review of 10 genes (Additional file 2: Figure S3A). Furthermore, VAAST and Exomiser failed to provide rankings for 4 and 18 true positive variants, respectively. Exclusion of false negatives and SV cases increased the top 10 recall of Exomiser to 93% (Additional file 2: Figure S3B), in agreement with previous reports [86]. These data show the importance of including all types of cases and causal variants in benchmarking to avoid overestimation of diagnostic performance in real-world clinical applications.

Scoring of structural variants increases diagnostic rate

A major barrier to the incorporation of SV calls into genome diagnostic interpretation, whether manual or using eCDSS, is their low precision (high false positive, FP, rates) using short read alignments, with typical FP rates of 20–30% [87, 88]. This leads to overwhelmingly time-consuming, manual assessment of event quality and significance for large numbers of SVs. GEM minimizes the effect of low precision by scoring SVs either with SV calls provided in the proband's input VCF file, and/or by inferring ab initio their existence from metadata associated with SNV and indel calls ("Methods"; see below). The benchmark cohort included 20 cases in which SVs were reported to be causative, reflecting a similar incidence to that in real-world experience (Fig. 1A, Table 2) [20–23]. In 17 of these, the causative SV was ranked first by GEM. In two, it was ranked second, and in one it was listed fourth, demonstrating that GEM retains adequate diagnostic performance with imprecise SV calls. The disease-causing SVs in the benchmark set ranged from small (4 kb) to very large (e.g., entire chromosome arms). In three cases, the diagnosis was of an autosomal recessive disorder in which the SV was compound heterozygous with a SNV/indel. In each, GEM integrated the two variants correctly, automatically identifying the causative diplotypes (Additional file 2: Figure S5). With regard to the diagnostic specificity of GEM, the mean and median number of gene candidates for these probands with BF > 0 (any support) was 8.7 and 9.5, respectively, which was similar to probands whose VCF files contained no SVs, causative or otherwise.

Large SVs frequently affect more than one gene. For consistency with other variant classes, genes within multigenic SVs are grouped and sorted by GEM based upon the gene-centric Bayes factor score associated with the overlap of the proband phenotype and known Mendelian disorders (if any) associated with them ("Methods"). For example, Additional file 2: Figure S4 shows a case that highlights the practical utility of prioritizing genes harboring causative SVs together with SNVs and short indels in the same report, rather than separately cross-referencing with databases of microdeletion syndromes [89]. While it is often unknown which genes harbored in a pathogenic SV are causal for microdeletion/microduplication syndromes, GEM's gene-by-gene rankings typically agreed with causal gene candidates suggested by the literature (asterisks in Table 2).

By default, GEM evaluates every gene and transcript for the presence of overlapping SVs. Notably, four benchmark cases did not include externally called SVs in their input VCFs (these had been previously diagnosed by manual inspection and orthogonal confirmatory tests; Table 2). Nevertheless, GEM inferred the existence of these four SVs using its ab initio SV identification algorithm and evaluated them jointly with SNVs and indels ("Methods"). To further demonstrate this innovative functionality, we removed all external SV calls from each input VCF file of the 14 WGS cases (as GEMs ab initio SV imputation is currently limited to WGS data) and reran GEM. GEM re-identified 13 of the 14 of the causative SVs. Although GEM's inferred SV termini were imprecise, an overlapping SV of the same class (duplication, deletion, or CNV) and ploidy to that in the original VCF was inferred, and the same high scoring gene and mode of inheritance/genotype (autosomal dominant, simple recessive, or compound heterozygote) was ranked first. SV recall within the top 1, 5, and 10 ranked GEM results were 71%, 86%, and 93%, respectively. The single false negative was a small (4 kb) homozygous deletion. GEM failed to identify this SV because it did not span sites with known variation in the gnomAD database [73], upon which ab initio SV inference is based ("Methods"). With regard to specificity, the mean and median number of results with genes with BF > 0 in these cases was 10.6 and 12.5, respectively. These values differed only slightly from the results obtained using external SV calls (8.7 and 9.5, respectively), despite the fact every gene and transcript was evaluated for the presence of SVs.

Collectively, these results demonstrate the accuracy of GEM's ab initio approach to identification and prioritization of SVs without recourse to external calls and databases of known causative SVs. Thus, GEM compensates, in part, for the low recall of SVs from short-read sequences. If an external SV calling pipeline fails to detect an SV, there is still the possibility that GEM will identify it via this ab initio approach. This capability, together with GEM's ability to accurately prioritize SVs in the context of SNVs and short indels, addresses an unmet need for clinical applications. This characteristic also makes GEM well suited for reanalyses of older cases and/or pipelines lacking SV calling.

Leveraging automated phenotyping from clinical natural language processing

Ontology-based phenotype descriptions, using Human Phenotype Ontology (HPO) terms [69], are widely used to communicate the observed clinical features of disease in a machine-readable format. These lists of terms are usually derived by manual review of patient EHR data by trained personnel, a time-consuming, subjective process. A solution is automatic extraction of patient phenotypes from clinical notes using clinical natural language processing (CNLP) [28, 90]. One challenge has been that CNLP generates many more terms than manual extraction. Thus, manual curation yielded an average of 4 HPO terms (min = 1, max = 12) in the benchmark cohort, while CNLP yielded an average of 177 HPO terms (min = 2, max = 684). Some of the extra CNLP terms are hierarchical parent terms of those observed, raising the concern that their inclusion diminishes the average information content in a manner that could impede diagnosis [27]. To investigate the effect of CNLP-derived HPO terms on GEM's performance, we analyzed the benchmark cohort both with HPO terms extracted by commercial CNLP ("Methods") and manually extracted HPO terms.

Figure 2 shows the distributions and medians for ranks and GEM gene scores of true positives, as well as the number of gene candidates with BF ≥ 0.69 (moderate support), for manual and CNLP terms. The median rank of the causal genes did not significantly differ between CNLP- and manually derived phenotype descriptions (Fig. 2A). The median GEM gene score of true positives was higher with CNLP-derived phenotypes than with manual phenotypes (Fig. 2B). The number of candidates above the BF threshold was higher with manual phenotypes than CNLP (Fig. 2C). CNLP rescued a few true positives with low ranks and negative BF scores compared to manual phenotype descriptions (Fig. 2A, B). These results demonstrate that GEM performs somewhat better with CNLP-derived phenotype descriptions as part of an automated interpretation workflow, than with sparse, manual phenotypes.

Resilience to mis-phenotyping and gaps in clinical knowledge

Given the potentially noisy nature of the CNLP phenotype descriptions, it was important to examine the sensitivity of GEM to mis-phenotyping. To address this question, we randomly permuted CNLP-extracted HPO terms between cases, weighting by term frequency within the cohort, so that every case maintained the same number of HPO terms as CNLP originally provided. Permuting HPO terms resulted in lower gene scores, and several cases would have been lost for review had the gene score threshold of BF ≥ 0 still been used,

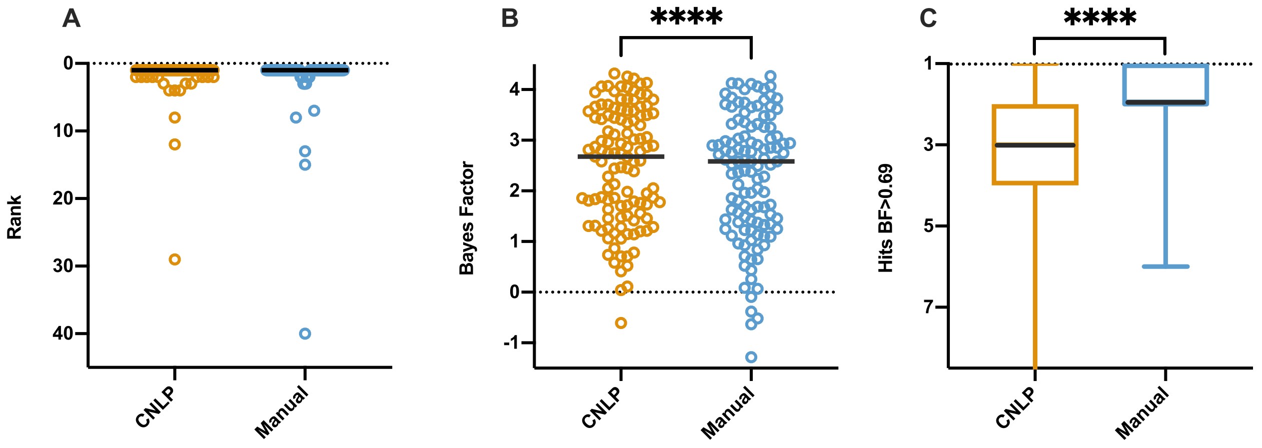


Fig. 2 Comparison of GEM performance with manually curated and CNLP-derived HPO terms in the benchmark cohort. Distribution of ranks for causal genes (**A**); GEM Bayes factors for causal genes (**B**); and number of candidates (hits) at BF ≥ 0.69 threshold (moderate support) (**C**). The black line in the graphs denotes the median. The asterisks represent statistical difference between the groups with $p < 0.0001$ from a two-tailed Wilcoxson matched pairs signed rank test (ranks showed no statistically significant difference)

but ranks are unaffected (98% in top 10; Fig. 3). This represented lower bound estimates, as actual mis-phenotyping (short of data tracking issues) would be much less. It is also worth noting that even using randomly permuted phenotype descriptions, GEM's performance still exceeded that of Phevor and Exomiser using the correct phenotypes (Additional file 2: Figure S2). We therefore conclude that GEM is resilient to mis-phenotyping.

We also evaluated the impact of gaps in clinical knowledge on GEM performance by withdrawing annotations from a key clinical database, ClinVar. Absence of ClinVar annotations had minimal impact in ranking, although it reduced median gene scores (1.1 vs. 2.7), resulting in 9 cases no longer meeting the minimum Bayes factor threshold ≥ 0 (any support; Fig. 3). Clearly, ClinVar provided GEM with valuable information. Nonetheless, without ClinVar, GEM's top 10 maximal recall (88%) still exceeded that of Phevor (72%) and Exomiser (65%; Fig. 1). More broadly, these results show that integrating more datatypes in GEM improves diagnostic performance and results in greater algorithmic stability (Figs. 2 and 3).

About 70% (86/122) of the disease-causing variants in the benchmarking dataset are reported in ClinVar with pathogenic (P) or likely pathogenic (LP) clinical significance annotations. Moreover, each proband's whole-genome variant set contained on average 1.9 variants

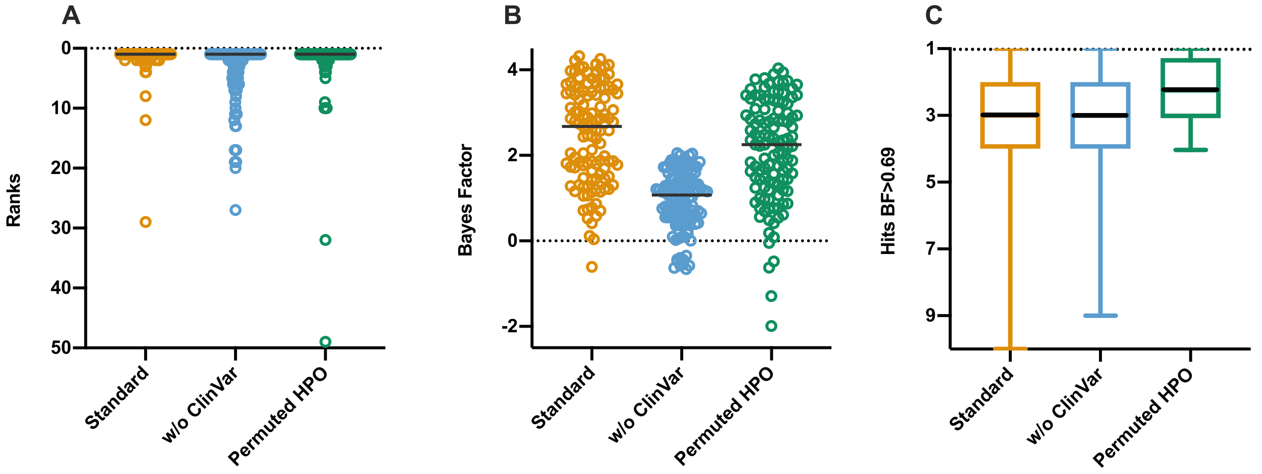


Fig. 3 Impact of missing data and mis-phenotyping on GEM performance in the benchmark cohort. Causal gene rank (**A**); Bayes factors for causal genes (**B**); and number of candidates (hits) above gene BF ≥ 0.69 threshold (moderate support) (**C**) under standard conditions, withdrawing ClinVar information, and permuting HPO terms extracted by CNLP. The black line in the graphs denotes the median

with ClinVar P/LP annotations. These two facts underscore the importance of ClinVar annotations for assisting diagnosis. They also make clear that tools that leverage ClinVar information need to avoid false positives which lead to longer candidate lists as non-causal genes also contain ClinVar P/LP variants. Additional file 1: Table S4 breaks down results for the benchmark cohort with respect to ClinVar annotations of causal variants. Overall, mean, and median ranks were slightly improved for diagnostic variants with ClinVar P/LP annotations *vs.* those without them (mean 1 *vs.* 3), with GEM showing the greatest improvement in ranks. Moreover, GEM maintained the same number of candidates with GEM gene score > 0 for both classes [10], demonstrating that GEM can use ClinVar status to improve diagnostic rates without increasing the number of candidates for review.

GEM performs equivalently on parent-offspring trios and single probands

Parent-offspring trios are widely used for molecular diagnosis of rare genetic disease. While a recent study showed that singleton proband sequencing returned a similar diagnostic yield as trios [91], interpretation of trio sequences is less labor-intensive. For example, trios enable facile identification of de novo variants, which is the leading mechanism of genetic disease in outbred populations [92]. Likewise, in recessive disorders, proband compound heterozygosity can be automatically distinguished from two variants in *cis*. However, these benefits are associated with increased sequencing costs. Moreover, both parents are not always available for sequencing or do not wish to have their genomes sequenced.

To understand how GEM performs in the absence of parental data, we reanalyzed the 63 trio and duo cases from the benchmark cohort as singleton proband cases. Surprisingly, we observed insignificant differences in the mean rank of the causal gene (Fig. 4A), GEM score of the causal gene (Fig. 4B), or number of candidates with BF ≥ 0.69 (Fig. 4C), using either manually or CNLP-extracted HPO terms. In contrast, this reanalysis was associated with a decline in the performance of Exomiser (Additional file 2: Figure S6). These analyses demonstrated that GEM was resilient to the absence of parental genotypes, a feature that could increase the cost effectiveness and adoption of WGS.

GEM scores optimize case review workflows

Conventional prioritization algorithms rank variants, enabling manual reviewers to start with the top ranked variants, and work their way down in the list until a convincing variant is identified for further curation, classification, and possible clinical reporting. This review process typically involves (a) assessing variant quality, deleteriousness, and prior clinical annotations; (b) evaluating whether there is a reasonable match between the phenotypes exhibited by the patient and those reported for condition(s) known to be associated with defects in the corresponding gene; and (c) considering the match in mode(s) of inheritance reported in the literature for the candidate disease and the patient's diplotype.

GEM accelerates this process, because it intrinsically considers variant quality, deleteriousness, prior clinical annotations, and mode of inheritance. Furthermore, at manual review, GEM gene scores summarize the relative strength of evidence supporting the hypothesis that the

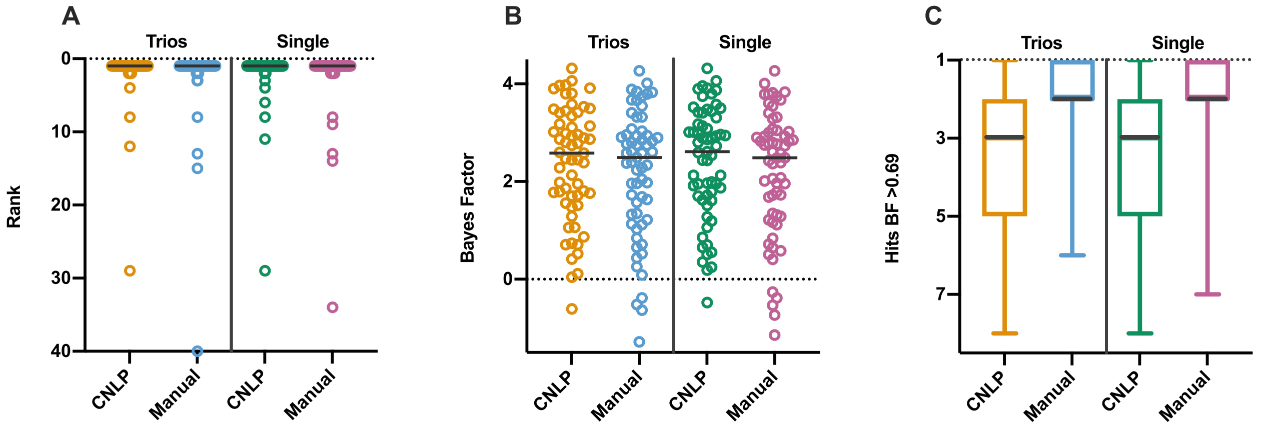


Fig. 4 Comparative performance of parent-offspring trios or duos vs. singleton probands in the benchmark cohort. Causal gene rank (**A**); Bayes factors (**B**); and number of candidates (hits) above gene BF ≥ 0.69 (moderate support) (**C**) for 63 cases analyzed as parent-offspring trios (*n* = 59) or duos (*n* = 4), as compared with analysis as single probands, using both manually curated or CNLP-derived HPO terms. The black line in the graphs denotes the median. No statistically significance difference between the any manual/CNLP groups was found between trios versus single probands using the two-tailed Wilcoxson matched pairs signed rank test

gene is damaged and that this explains the proband's phenotype.

GEM scores provide a logical framework for setting thresholds with regard to the optimal number of candidates that should be reviewed to achieve a desired diagnostic rate. This enables laboratory directors and clinicians to dynamically set optimal tradeoffs of interpretation time and diagnosis rate for specific patients, relative to their suspicion of a genetic etiology or results of other diagnostic tests.

We examined the effect of different BF thresholds on recall (true positive rate) and median number of gene candidates for review in the benchmark cohort (Fig. 5). In such analyses, it is germane to consider the concept of *maximal true positive rate (or recall)* to measure the theoretical proportion of true positive diagnoses recoverable by perfect interpretation when reviewing a set of *N* genes containing the true positive. For example, in the benchmark dataset, a GEM causal gene score threshold ≥ 0 would retain a median of ten candidates for review and provide a 99% maximal recall; whereas a threshold of ≥ 0.5 would retain a median of four candidates for review for a 97% maximal recall (Fig. 5).

These results illustrate how a tiered approach to case review using GEM gene scores could minimize the number of candidate genes to review, and, thereby manual interpretation effort. For example, a first pass review of candidates with a gene BF ≥ 0.69 provided an expected 95% diagnostic rate (and a corresponding median of 3 genes to be manually reviewed). If followed by a second pass using a threshold > 0, if no convincing candidates are found, an additional 4% possible diagnoses would be recovered, involving review of a median increment of seven genes. Application of this two-tiered approach to the benchmark dataset of 119 cases (Fig. 1), required manual final review of 395 candidate genes (3 genes in 115 cases and 10 genes in 5 cases), or an average of 3.3 candidate genes per case, for a maximal recall of 99%. Finally, review of candidates with BF < 0 recovered the last true positive in the benchmark cohort (*COL4A4*, ranked 40th in the GEM report with a BF = − 0.6. This case was a phenotypically and genotypically atypical autosomal dominant presentation of Alport syndrome 2 (MIM 203780).

Clinical decision support for diagnosis

Quantifying how well the observed phenotypes in a patient match the expected phenotypes of Mendelian conditions associated with a candidate gene is challenging for clinical reviewers and is a major interpretation bottleneck. In practice, clinicians look for patterns of phenotypes, biasing their observations. In addition, patient phenotypes evolve as their disease progresses. And there is considerable, disease-specific heterogeneity in the range of expected phenotypes. Simply comparing exact matches of the patient's observed HPO terms with those expected for that disease is suboptimal, because the observed and expected HPO terms are often hierarchical neighbors, rather than exact matches. Missing terms, particularly those considered pathognomonic for a condition, and "contradictory" terms further complicate such comparisons by clinicians. Thus, generation of quantitative, standardized, unbiased models of disease similarity has proven elusive.

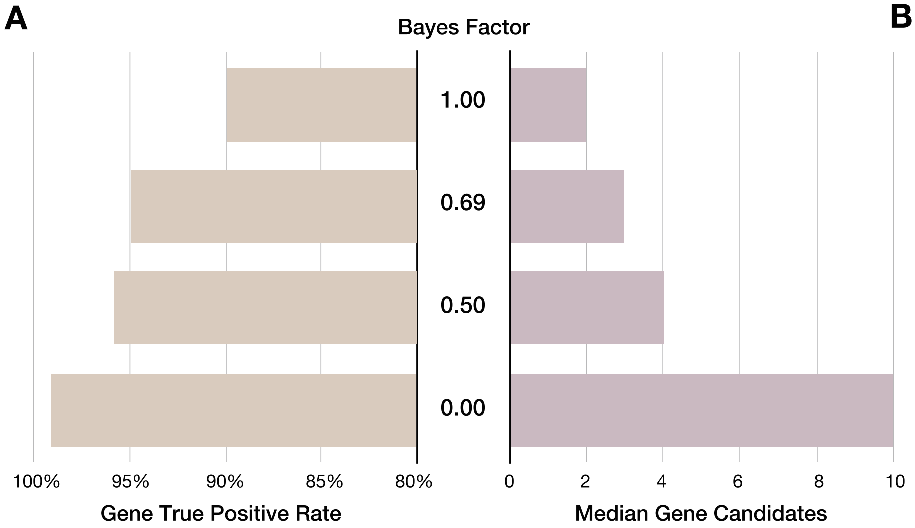


Fig. 5 Trade-off between GEM gene scores, maximal true positive rates, and number of candidates for review in the benchmark cohort. GEM gene scores are Bayes factors (BF) that can be used speed case review. **A** Gene maximal true positive rate achieved at the different BF thresholds (*Y*-axis). **B** Median number of candidate genes for review at each BF threshold. As the BF threshold is decreased, true positive rate increases, while the number of candidates to review remains manageable. Input HPO terms for this analysis were extracted by CNLP

GEM can automate or provide clinical decision support for this process via a condition match (CM) score ("Methods"). The GEM CM score summarizes the match between observed and expected HPO phenotypes for genetic diseases and considers the known mode(s) of inheritance, associated gene(s), their genome location(s), proband sex, the pathogenicity of observed diplotypes, and ClinVar annotations. Importantly, CM scores reflect relationships between phenotype terms as expressed in the HPO ontology graph, enabling inclusion of imprecise matches in similarity comparisons. CM scores can be used in a wide variety of clinical settings to prioritize and quickly assess possible Mendelian conditions as candidate diagnoses, a process we term *diagnostic nomination*.

Specific, definitive, genetic disease diagnosis remains a significant challenge for clinical reviewers, even with the short, highly informative candidate gene lists provided by tools such as GEM. This is because many genes are associated with more than one Mendelian disease. For example, application of a GEM causal gene score threshold ≥ 0.69 to the 119 probands in the benchmark cohort results in a median of 3 gene candidates per proband (c.f. Fig. 5), associated with a maximal gene recall of 95%. However, because many genes are associated with more than one disease, clinical reviewers would actually need to consider around 12 candidate Mendelian conditions per proband (data not shown). This difficulty is exacerbated by the fact that most laboratory directors are not physicians and lack formal training in clinical diagnosis.

Determination of a specific, definitive genetic disease diagnosis among several candidates can be accomplished with a combination of GEM CM scores and causal gene scores (Fig. 6). Using the benchmark cohort's true (reported) gene and disorder diagnoses as ground truth, we used a GEM gene score threshold ≥ 0.69 to recover gene candidates, and the associated CM scores to rank order the diseases associated with those gene candidates (Fig. 6A). Using CNLP-derived phenotypes, the true disease diagnosis was the top nomination by CM score in 75% of cases, within the top 5 in 91% of cases, and within the top 10 in 95% of cases. Performance was inferior with manually extracted phenotype terms. The area under the receiver-operator characteristic (ROC) curves (AUCs) were 0.90 and 0.88, for CNLP and manual terms, respectively (Fig. 6B). This implied that the larger number of CNLP-extracted terms conveyed greater information content, permitting better discrimination of the correct diagnostic condition, than sparse, manually extracted phenotypes [27].

In the benchmark cohort, 58 of the 100 candidate genes (excluding cases with causal, multigenic SVs) were associated with 2 or more disorders (median of 3 gene-disorder, maximum of 15; Additional file 2: Figure S7 shows the example of *ERCC6*). We measured how well the CM score distinguished between multiple, alternative disorders associated with the same gene (Fig. 6B). In these 58 cases, the AUC was less than that for CNLP with the entire set of candidate genes in the benchmark cohort (0.68 vs 0.9). This decrease can be at least partially explained by the high similarity (and in some cases identity) of the clinical features of different disorders associated with the same gene. Thus, a combination of GEM gene and CM scores can refine candidate disorders for clinical reporting, further reducing review times.

Reanalysis of previously unsolved cases

Recent reports show that reanalysis of older unsolved cases suspected of rare genetic disease can yield new

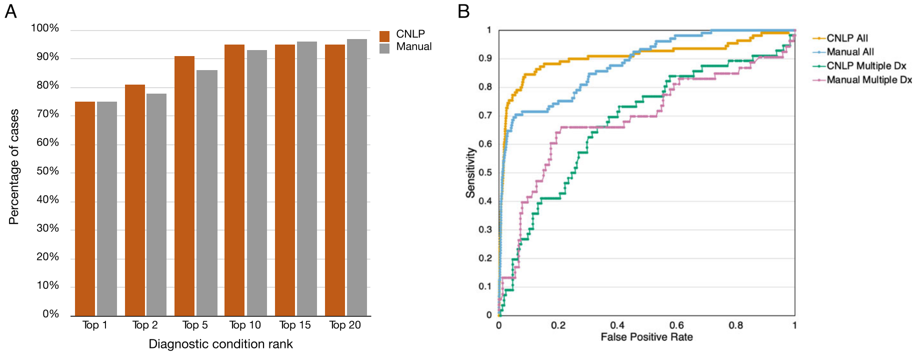


Fig. 6 Performance of GEM condition match scores for diagnostic nomination in the benchmark cohort. **A** Ranks for reported diagnostic conditions for the benchmark dataset, using a GEM gene BF score ≥ 0.69 and sorted by CM score, for HPO terms derived from CNLP or manual curation. **B** Receiver-operator characteristic curves for the condition match **(**CM) score for all hits with BF ≥ 0. CNLP All: HPO extracted from clinical notes by CNLP; AUC = 0.91. Manual: Manually curated HPO terms; AUC = 0.88. CNLP Multiple Dx: CNLP-derived CM score for the true positive disorder versus the other possible disorders associated with that gene; AUC = 0.68. Manual Multiple Dx: As for CNLP-derived CM but using manually curated HPO terms; AUC = 0.69

diagnoses supported by incremental increases in knowledge of pathogenic variants, disease-gene discoveries, and reports of phenotype expansion for known disorders [93, 94]. While worthwhile, there are barriers to reanalysis, such as limited reimbursement and low incremental diagnostic yield, that limit use to physician requests. Ideally, all unsolved cases would be reanalyzed automatically periodically, and a subset with high likelihood of new findings would be prioritized for manual review. The strong correlation between true positive rates and GEM gene scores (Fig. 5) suggested a strategy for triaging reanalyzed cases for manual review: only cases for which the recalculated GEM score had increased sufficiently to suggest a high probability of a new diagnosis would pass the threshold for manual review. Likewise, GEM condition match scores could be used to search all prior cases to identify the subset of unsolved cases with support for particular Mendelian conditions, aiding cohort assembly for targeted reanalysis based upon particular proband phenotypes, or for review by particular medical specialists. Of note, an advantage of CNLP is that it is possible to automatically generate a new clinical feature list at time of reanalysis. This is particularly important in disorders whose clinical features evolve with time and were the observed features may be nondescript at presentation.

To test the utility of GEM for reanalysis, we selected 14 unsolved cases that had rWGS performed by RCIGM. For these reanalyzes, we used CNLP-derived HPO terms (Table 3) and a more stringent gene BF threshold ≥ 1.5 to restrict the search to very strongly supported candidates. Ten cases yielded no hits. Four cases returned a total of 7 candidate genes. Review of three cases did not return new diagnoses. In the remaining case, a new likely diagnosis was made of autosomal dominant Shwachman-Diamond Syndrome (MIM: 260400) or severe congenital neutropenia (MIM: 618752) [95, 96], both of which are associated with pathogenic variants in *SRP54*. The respective CM scores using 261 CNLP-derived terms were relatively high (0.893 and 0.672, respectively). The association of *SRP54* and these disorders was first reported in November 2017 [95] and entered in OMIM in January 2020 [97], which explained why it was not identified as the diagnosis originally in July 2017. The identified candidate p.Gly108-Glu variant has been classified as “uncertain significance” by ACMG guidelines. However, if we were able to confirm de novo origin with paternal genotypes (which is currently lacking for this single proband case), the variant could be reclassified as “likely pathogenic” (meeting PM2, PM1, PP3, and PM6 of the ACMG guidelines). This was a singleton proband sequence and confirmation is being pursued. Thus, GEM reanalysis of 14 unsolved cases led to 7 gene-disorder reviews (an average of 0.5 per case), and yielded one likely new diagnosis, which was consistent with prior reanalysis yields [93, 94].

Conclusions

Here we described and benchmarked a Bayesian, AI-based gene prioritization tool for scalable diagnosis of rare genetic diseases by CNLP and WES or WGS. GEM improved upon prior, similar tools [19, 27, 28, 98, 99] by incorporating OMIM, HPO, and ClinVar knowledge explicitly, automatically controlling for confounding factors, such as sex and ancestry, compatibility with CNLP-derived phenotypes, SVs and singleton probands, and by directly nominating diplotypes and disorders, rather than just prioritizing variants.

In the cohorts examined, GEM had maximal recall of 99%, requiring review of an average of 3 candidate genes, and less than one half of the associated disorders nominated by other widely used variant prioritization methods per case. Improved diagnostic performance is anticipated to enable faster and more cost-effective, tiered reviews. GEM recall was essentially unaltered in the absence of parental genotypes in our data, meaning that full trio-sequencing is not always a requirement for high diagnostic yield. However, our cohort includes only definitively solved cases with 70% of variants already

Table 3 Previously undiagnosed cases with potential leads. Cases with hits with a GEM gene score BF > 1.5. *Zygo* zygosity, *Hom* homozygous, *Het* heterozygous, *Dup* large duplication

Case	Pedigree	Sex	Assay	Rank	Chr	Gene	Variant Type	Variant ACMG	De novo	Zygo	GEM score	Mode of inheritance	MIM ID(s)	CM score(s)
244799	Single	Male	WGS	1	14	SRP54	SNV	Uncertain significance	Likely	Het	1.76	Dominant	618752, 260400	0.672, 0.893
245237	Trio	Male	WGS	2	X	GK	SNV	VUS	Yes	Het	1.60	X-linked recessive	307030	1.119
245237	Trio	Male	WGS	3	16	FANCA	SNV	VUS	No	Hom	1.55	Recessive	227650	1.315
245768	Single	Male	WGS	1	16	TSC2	Dup	VUS	Likely	Het	1.64	Dominant	N/A	N/A
247458	Single	Male	WGS	1	1	SLC25A24	SNV	VUS	Likely	Het	1.86	Dominant	612289	1.995
247963	Trio	Female	WGS	1	X	STAG2	SNV	Likely pathogenic	Yes	Het	1.53	X-linked dominant	301022	1.25

classified as P/LP in ClinVar; identification of less certain candidate variants and genes may still benefit from parental genotypes for ascertaining de novo variants, and for phasing alleles in genes associated to recessive conditions.

Uniquely, GEM provided AI-based unified gene prioritization for SVs and small variants. Hitherto, this has been frustrated by the high false positive rates of SV calls using short-read sequences and lack of a suitable framework for AI-based SV pathogenicity assertions [87, 88]. Furthermore, GEM inferred SV calls ab initio from WGS when they were not provided. These functionalities are critical for reanalyzing older cases, and for pipelines lacking SV calls.

Finally, in a small data set, we showed that GEM can efficiently reanalyze cases, potentially permitting cost-effective, scalable reanalysis of previously unsolved cases as disease, gene, and variant knowledge evolves [94, 100]. Indeed, integration of GEM and CNLP could enable automatic surveillance for rare disease patients [101] from genomes obtained for research or other clinical tests performed in healthcare [102, 103]. These combined features hold promise for reduced time-to-diagnosis and greater scalability for critical applications, such as in seriously ill children in the NICU/PICU [27, 104].

Abbreviations

AI: Artificial intelligence; BF: Bayes factor; eCDS: Electronic Clinical Support System; CNLP: Clinical Natural Language Processing; CM: Condition match score; GEM: Fabric GEM; HPO: Human Phenotype Ontology; OMIM: Online Mendelian Inheritance in Man; SNV: Single-nucleotide variant; Indel: Insertion-deletion; SV: Structural variants; VAAST: Variant Annotation, Analysis and Search Tool; VVP: VAAST Variant Prioritizer; WES: Whole-exome sequencing; WGS: Whole-genome sequencing

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13073-021-00965-0.

Additional file 1. Supplementary Tables (Tables S1-S14).

Additional file 2. Supplementary Figures (Figures S1-S7).

Acknowledgements

We thank Jeff Rule and Birgit Crain for help in extracting case data for the RCH cases from Fabric Enterprise, Brent Lutz for project management, and Sandy White for interinstitutional coordination (Fabric Genomics, Oakland CA). We are grateful to Joe Azure, Josh Grigonis, Jeff Rule, Bjoern Achilles-Stade, Peter Spiro, Gilberto De La Vega, Tara Friedrich (Fabric Genomics, Oakland, CA), Ray Drummond, Richard Littin, Aidan Scarlet, Mike Richdale, and Tim Dawson (NetValue Ltd, Hamilton, New Zealand) for software and architecture development efforts to implement GEM in the Fabric Enterprise platform. We also acknowledge Edward Kiruluta and Marco Falcioni (formerly Fabric Genomics) for useful discussions early on the project.

Authors' contributions

FV, MY, and SC designed the study and analysis strategy. MY developed the GEM algorithms. FV guided requirements, designed UIs, and led the software implementation. BM and EF implemented analysis pipelines. BM, EF, EJH, JM, FV, and MY performed data analysis. AF, AHB, BB, BSL, CAB, CAG, JM, KJ, KÕ, KR, LG, MH, MN NV, PBA, SC, SP, TW, and VN complied cases and clinical evidence. PB provided feedback on features and development. MGR and SK sponsored the project and provided helpful discussions and edits of the manuscript. FV and MY wrote the manuscript. All authors reviewed and suggested edits for the final version of the manuscript. The authors read and approved the final manuscript.

Funding

MH, KR, MN, and VN were supported in part by The Center for Rare Childhood Disorders, funded through donations made to the TGen Foundation. AF and BSL were supported by the DFG Cluster of Excellence "Precision Medicine in Chronic Inflammation". KO and SP were supported by Estonian Research Council grants PUT355, PRG471, MOBTP175, and PUTJD827. Sequencing and analysis were partially provided by the Broad Institute of MIT and Broad Center for Mendelian Genomics (Broad CMG) and was funded by the National Human Genome Research Institute, the National Eye Institute, and the National Heart, Lung and Blood Institute grant UM1 HG008900 and in part by National Human Genome Research Institute grant R01 HG009141. The phenotyping and analysis of patients at Boston Children's Hospital was funded by MDA602235 from the Muscular Dystrophy Association, and the Tommy Fuss Foundation, and the Yale Center for Mendelian Genomics. Sanger sequencing confirmations utilized the resources of the Boston Children's Hospital IDDRC Molecular Genetics Core Facility supported by U54HD090255 from the National Institutes of Health.

Availability of data and materials

The datasets supporting the conclusions of this article are included within the article and its additional files. Due to patient privacy, data sharing consent, and HIPAA regulations, our raw data cannot be submitted to publicly available databases. Anonymized outputs from GEM [70], Phevor [15], VAAST [14], and Exomiser [16] for the benchmark dataset cases are tabulated in Additional file 1: Tables S5-S8, and GEM for the validation data-set cases in Additional file 1: Table S10. Condition match scores for hits with gene BF > 0 used for Fig. 6 are tabulated in Additional file 1: Tables S11-S14. GEM, Phevor, and VAAST software implementations for versions used in this analysis are part of the Fabric Enterprise analysis platform and are commercially available [70]. Exomiser source code (version 12.1.0) is available on GitHub [105].

Declarations

Ethics approval and consent to participate

The need for Institutional Review Board Approval at Rady Children's Hospital for the current study was waived as all data used from this project had previously been generated as part of IRB approved studies and none of the results reported in this manuscript can be used to identify individual patients. The studies from which cases derived from were previously approved by the Institutional Review Boards of Rady Children's Hospital, Boston Children's Hospital (IRB protocols 03-08-128R and 10-02-0053), Christian-Albrechts University of Kiel (approval #A-156/02), HudsonAlpha Institute for Biotechnology (Western Institutional Review Board #20130675 and the University of Alabama at Birmingham #X130201001), the Translational Genomics Research Institute (WIRB® Protocol #20120789), and the Research Ethics Committee of the University of Tartu (approvals #263/M-16 and #2871N). These studies were performed in accordance with the Declaration of Helsinki and informed consent was obtained from at least one parent or guardian.

Consent for publication

Not applicable.

Competing interests

FV, EF, JM, and MGR were employees of Fabric Genomics Inc. during the performance of this work and have received stock grants from Fabric Genomics Inc. BM, PB, and MY are consultants to Fabric Genomics Inc. and have received consulting fees and stock grants from Fabric Genomics Inc. The remaining authors declare that they have no competing interests.

Author details

[1]Fabric Genomics Inc., Oakland, CA, USA. [2]Department of Biomedical Data Science, Stanford University School of Medicine, Stanford, CA, USA. [3]Current Address: Tempus Labs Inc., Redwood City, CA 94065, USA. [4]Rady Children's

Institute for Genomic Medicine, San Diego, CA, USA. [5]Department of Human Genetics, Utah Center for Genetic Discovery, University of Utah, Salt Lake City, UT, USA. [6]Division of Genetics and Genomics, The Manton Center for Orphan Disease Research, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA. [7]Division of Newborn Medicine, Boston Children's Hospital, Boston, MA, USA. [8]Institute of Clinical Molecular Biology, Christian-Albrechts-University of Kiel & University Hospital Schleswig-Holstein, Kiel, Germany. [9]HudsonAlpha Institute for Biotechnology, Huntsville, AL, USA. [10]Department of Clinical Genetics, United Laboratories, Tartu University Hospital, Tartu, Estonia. [11]Department of Clinical Genetics, Institute of Clinical Medicine, University of Tartu, Tartu, Estonia. [12]Center for Rare Childhood Disorders, Translational Genomics Research Institute, Phoenix, AZ, USA.

Received: 22 March 2021 Accepted: 27 August 2021
Published online: 14 October 2021

References

1. Church G. Compelling reasons for repairing human germlines. New Engl J Med. 2017;377:1909–11.
2. Bamshad MJ, Nickerson DA, Chong JX. Mendelian gene discovery: fast and furious with no end in sight. Am J Hum Genet. 2019;105:448–55.
3. Online Mendelian Inheritance in Man, OMIM®McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD) (available at https://omim.org/).
4. Wright CF, FitzPatrick DR, Firth HV. Paediatric genomics: diagnosing rare disease in children. Nat Rev Genet. 2018;10:1–16.
5. Mardis ER. The $1,000 genome, the $100,000 analysis? Genome Med. 2010;2:84.
6. Lavenier D, Cimadomo R, Jodin R. Variant calling parallelization on processor-in-memory architecture. bioRxiv 2020.11.03.366237.
7. Lee S, Min H, Yoon S. Will solid-state drives accelerate your bioinformatics? In-depth profiling, performance analysis and beyond. Brief Bioinform. 2015;17:713–27.
8. Kiely B, Vettam S, Adesman A. Utilization of genetic testing among children with developmental disabilities in the United States. Appl Clin Genet. 2016;9:93–100.
9. Markello TC, Adams DR. Current protocols in human genetics. Curr Protoc Hum Genet Éditor Board Jonathan L Haines Et Al. 2013;79:6.13.1–6.13.19.
10. Dewey FE, Grove ME, Pan C, Goldstein BA, Bernstein JA, Chaib H, et al. Clinical interpretation and implications of whole-genome sequencing. Jama. 2014;311:1035–45.
11. Jagadeesh KA, Wenger AM, Berger MJ, Guturu H, Stenson PD, Cooper DN, et al. M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. Nat Genet. 2016;48:1–8.
12. Yang Y, Muzny DM, Reid JG, Bainbridge MN, Willis A, Ward PA, et al. Clinical whole-exome sequencing for the diagnosis of mendelian disorders. New Engl J Med. 2013;369:1502–11.
13. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38:e164.
14. Yandell M, Huff C, Hu H, Singleton M, Moore B, Xing J, et al. A probabilistic disease-gene finder for personal genomes. Genome Res. 2011;21:1529–42.
15. Singleton MV, Guthery SL, Voelkerding KV, Chen K, Kennedy B, Margraf RL, et al. Phevor combines multiple biomedical ontologies for accurate identification of disease-causing alleles in single individuals and small nuclear families. Am J Hum Genet. 2014;94:599–610.
16. Robinson P, Kohler S, Oellrich A, Project SMG, Wang K, Mungall C, et al. Improved exome prioritization of disease genes through cross species phenotype comparison. Genome Res. 2013;24. https://doi.org/10.1101/gr.160325.113.
17. Agrawal S, Javed A, Ng PC. Phen-Gen: combining phenotype and genotype to analyze rare disorders. Nat Methods. 2014;11:1–7.
18. Yang H, Robinson PN, Wang K. Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. Nat Methods. 2015;12:841–3.
19. Birgmeier J, Haeussler M, Deisseroth CA, Steinberg EH, Jagadeesh KA, Ratner AJ, et al. AMELIE speeds Mendelian diagnosis by matching patient phenotype and genotype to primary literature. Sci Transl Med. 2020;12:eaau9113.
20. Cooper GM, Coe BP, Girirajan S, Rosenfeld JA, Vu TH, Baker C, et al. A copy number variation morbidity map of developmental delay. Nat Genet. 2011;43:838–46.
21. Miller DT, Adam MP, Aradhya S, Biesecker LG, Brothman AR, Carter NP, et al. Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. Am J Hum Genet. 2010;86:749–64.
22. Clark MM, Stark Z, Farnaes L, Tan TY, White SM, Dimmock D, et al. Meta-analysis of the diagnostic and clinical utility of genome and exome sequencing and chromosomal microarray in children with suspected genetic diseases. Npj Genom Med. 2018;3:1–10.
23. Yuan H, Shangguan S, Li Z, Luo J, Su J, Yao R, et al. CNV profiles of Chinese pediatric patients with developmental disorders. Genet Med. 2021:1–10.
24. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25:1–13.
25. Dias R, Torkamani A. Artificial intelligence in clinical and genomic diagnostics. Genome Med. 2019;11:70.
26. Birgmeier J, Deisseroth CA, Hayward LE, Galhardo LMT, Tierno AP, Jagadeesh KA, et al. AVADA: toward automated pathogenic variant evidence retrieval directly from the full-text literature. Genet Med. 2020;22:362–70.
27. Clark MM, Hildreth A, Batalov S, Ding Y, Chowdhury S, Watkins K, et al. Diagnosis of genetic diseases in seriously ill children by rapid whole-genome sequencing and automated phenotyping and interpretation. Sci Transl Med. 2019;11:eaat6177.
28. James KN, Clark MM, Camp B, Kint C, Schols P, Batalov S, et al. Partially automated whole-genome sequencing reanalysis of previously undiagnosed pediatric patients can efficiently yield new diagnoses. Npj Genom Med. 2020;5:33.
29. Shortliffe EH, Sepúlveda MJ. Clinical Decision Support in the Era of Artificial Intelligence. Jama. 2018;320:2199.
30. Kingsmore SF, Cakici JA, Clark MM, Gaughran M, Feddock M, Batalov S, et al. A randomized, controlled trial of the analytic and diagnostic performance of singleton and trio, rapid genome and exome sequencing in ill infants. Am J Hum Genet. 2019:1–17.
31. Rusell SJ, Norvig P. Artificial Intelligence: A Modern Approach. 4th ed. Hoboken: Prarson; 2020.
32. Eilbeck K, Quinlan A, Yandell M. Settling the score: variant prioritization and Mendelian disease. Nat Rev Genet. 2017;18:1–14.
33. Smedley D, Jacobsen JOB, Jäger M, Köhler S, Holtgrewe M, Schubach M, et al. Next-generation diagnostics and disease-gene discovery with the Exomiser. Nat Protoc. 2015;10:2004–15.
34. Robinson PN, Ravanmehr V, Jacobsen JOB, Danis D, Zhang XA, Carmody L, et al. Interpretable clinical genomics with a likelihood ratio paradigm. Medrxiv. 2020:2020.01.25.19014803.
35. Farnaes L, Hildreth A, Sweeney NM, Clark MM, Chowdhury S, Nahas S, et al. Rapid whole-genome sequencing decreases infant morbidity and cost of hospitalization. Npj Genom Med. 2018;3:1–8.
36. Farnaes L, Nahas SA, Chowdhury S, Nelson J, Batalov S, Dimmock DM, et al. Rapid whole-genome sequencing identifies a novel GABRA1 variant associated with West syndrome. Mol Case Stud. 2017;3:a001776.
37. Hildreth A, Wigby K, Chowdhury S, Nahas S, Barea J, Ordonez P, et al. Rapid whole-genome sequencing identifies a novel homozygous NPC1 variant associated with Niemann–Pick type C1 disease in a 7-week-old male with cholestasis. Mol Case Stud. 2017;3:a001966.
38. Sanford E, Watkins K, Nahas S, Gottschalk M, Coufal NG, Farnaes L, et al. Rapid whole-genome sequencing identifies a novel AIRE variant associated with autoimmune polyendocrine syndrome type 1. Mol Case Stud. 2018;4:a002485.
39. Sanford E, Farnaes L, Batalov S, Bainbridge M, Laubach S, Worthen HM, et al. Concomitant diagnosis of immune deficiency and Pseudomonas sepsis in a 19 month old with ecthyma gangrenosum by host whole-genome sequencing. Mol Case Stud. 2018;4:a003244.
40. Briggs B, James KN, Chowdhury S, Thornburg C, Farnaes L, Dimmock D, et al. Novel Factor XIII variant identified through whole-genome sequencing in a child with intracranial hemorrhage. Mol Case Stud. 2018;4:a003525.
41. Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics. 2016;32:1220–2.
42. Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res. 2011;21:974–84.
43. Ravenscroft G, Miyatake S, Lehtokari V-L, Todd EJ, Vornanen P, Yau KS, et al. Mutations in KLHL40 are a frequent cause of severe autosomal-recessive nemaline myopathy. Am J Hum Genet. 2013;93:6–18.

44. Konersman CG, Freyermuth F, Winder TL, Lawlor MW, Lagier-Tourenne C, Patel SB. Novel autosomal dominant TNNT1 mutation causing nemaline myopathy. Mol Genet Genom Med. 2017;5:678–91.
45. Lehtokari V, Kiiski K, Sandaradura SA, Laporte J, Repo P, Frey JA, et al. Mutation update: the spectra of nebulin variants and associated myopathies. Hum Mutat. 2014;35:1418–26.
46. Laing NG, Dye DE, Wallgren-Pettersson C, Richard G, Monnier N, Lillis S, et al. Mutations and polymorphisms of the skeletal muscle r-actin gene (ACTA1). Hum Mutat. 2009;30:1267–77.
47. Smedemark-Margulies N, Brownstein CA, Vargas S, Tembulkar SK, Towne MC, Shi J, et al. A novel de novo mutation in ATP1A3 and childhood-onset schizophrenia. Mol Case Stud. 2016;2:a001008.
48. Yuen M, Sandaradura SA, Dowling JJ, Kostyukova AS, Moroz N, Quinlan KG, et al. Leiomodin-3 dysfunction results in thin filament disorganization and nemaline myopathy. J Clin Invest. 2015;125:456–7.
49. D. D. D. (DDD) Study, Zweier M, Begemann A, McWalter K, Cho MT, Abela L, et al. Spatially clustering de novo variants in CYFIP2, encoding the cytoplasmic FMRP interacting protein 2, cause intellectual disability and seizures. Eur J Hum Genet. 2019;27:747–59.
50. Zeissig S, Petersen B-S, Tomczak M, Melum E, Huc-Claustre E, Dougan SK, et al. Early-onset Crohn's disease and autoimmunity associated with a variant in CTLA-4. Gut. 2015;64:1889.
51. Zeissig Y, Petersen B-S, Milutinovic S, Bosse E, Mayr G, Peuker K, et al. XIAP variants in male Crohn's disease. Gut. 2015;64:66.
52. Schubert D, Bode C, Kenefeck R, Hou TZ, Wing JB, Kennedy A, et al. Autosomal dominant immune dysregulation syndrome in humans with CTLA4 mutations. Nat Med. 2014;20:1410–6.
53. Müller T, Rasool I, Heinz-Erian P, Mildenberger E, Hülstrunk C, Müller A, et al. Congenital secretory diarrhoea caused by activating germline mutations in GUCY2C. Gut. 2016;65:1306.
54. Jung ES, Petersen B-S, Mayr G, Cheon JH, Kang Y, Lee SJ, et al. Compound heterozygous mutations in IL10RA combined with a complement factor properdin mutation in infantile-onset inflammatory bowel disease. Eur J Gastroen Hepat. 2018;30:1491–6.
55. Janecke AR, Heinz-Erian P, Yin J, Petersen B-S, Franke A, Lechner S, et al. Reduced sodium/proton exchanger NHE3 activity causes congenital sodium diarrhea. Hum Mol Genet. 2015;24:6614–23.
56. Amendola LM, Berg JS, Horowitz CR, Angelo F, Bensen JT, Biesecker BB, et al. The Clinical Sequencing Evidence-Generating Research Consortium: Integrating Genomic Sequencing in Diverse and Medically Underserved Populations. Am J Hum Genet. 2018;103:319–27.
57. Thompson ML, Finnila CR, Bowling KM, Brothers KB, Neu MB, Amaral MD, et al. Genomic sequencing identifies secondary findings in a cohort of parent study participants. Genet Med. 2018;20:1635–43.
58. East KM, Kelley WV, Cannon A, Cochran ME, Moss IP, May T, et al. A state-based approach to genomics for rare disease and population screening. Genet Med. 2021;23:777–81.
59. Bowling KM, Thompson ML, Amaral MD, Finnila CR, Hiatt SM, Engel KL, et al. Genomic diagnosis for children with intellectual disability and/or developmental delay. Genome Med. 2017;9:43.
60. Johnson BV, Kumar R, Oishi S, Alexander S, Kasherman M, Vega MS, et al. Partial loss of USP9X function leads to a male neurodevelopmental and behavioral disorder converging on transforming growth factor β signaling. Biol Psychiat. 2020;87:100–12.
61. Jepsen WM, Ramsey K, Szelinger S, Llaci L, Balak C, Belnap N, et al. Two additional males with X-linked, syndromic mental retardation carry de novo mutations in HNRNPH2. Clin Genet. 2019;96:183–5.
62. Puusepp S, Reinson K, Pajusalu S, Murumets Ü, Õiglane-Shlik E, Rein R, et al. Effectiveness of whole exome sequencing in unsolved patients with a clinical suspicion of a mitochondrial disorder in Estonia. Mol Genet Metab Rep. 2018;15:80–9.
63. Zimoń M, Baets J, Almeida-Souza L, Vriendt ED, Nikodinovic J, Parman Y, et al. Loss-of-function mutations in HINT1 cause axonal neuropathy with neuromyotonia. Nat Genet. 2012;44:1080–3.
64. Pravata VM, Gundogdu M, Bartual SG, Ferenbach AT, Stavridis M, Õunap K, et al. A missense mutation in the catalytic domain of O-GlcNAc transferase links perturbations in protein O-GlcNAcylation to X-linked intellectual disability. Febs Lett. 2020;594:717–27.
65. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl variant effect predictor. Genome Biol. 2016;17:1–14.
66. Flygare S, Hernandez EJ, Phan L, Moore B, Li M, Fejes A, et al. The VAAST Variant Prioritizer (VVP): ultrafast, easy to use whole genome variant prioritization tool. Bmc Bioinformatics. 2018;19:57.
67. Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, et al. ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res. 2016;44:D862–8.
68. Amberger JS, Bocchini CA, Scott AF, Hamosh A. OMIM.org: leveraging knowledge across phenotype–gene relationships. Nucleic Acids Res. 2018;47:gky1151.
69. Groza T, Köhler S, Moldenhauer D, Vasilevsky N, Baynam G, Zemojtel T, et al. The Human Phenotype Ontology: semantic unification of common and rare disease. Am J Hum Genet. 2015;97:111–24.
70. Fabric GEM (available at https://fabricgenomics.com/fabric-gem). Last accessed 22 Aug 2021.
71. 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491:56–65.
72. Sasani TA, Pedersen BS, Gao Z, Baird L, Przeworski M, Jorde LB, et al. Large, three-generation human families reveal post-zygotic mosaicism and variability in germline mutation accumulation. Elife. 2019;8:e46922.
73. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581:434–43.
74. Hoijtink H, Mulder J, van Lissa C, Gu X. A tutorial on testing hypotheses using the Bayes factor. Psychol Methods. 2019. https://doi.org/10.1037/met0000201.
75. Rabiner LR. A tutorial on hidden Markov models and selected applications in speech recognition. P Ieee. 1989;77:257–86.
76. Köster J, Rahmann S. Snakemake—a scalable bioinformatics workflow engine. Bioinformatics. 2012;28:2520–2.
77. Turro E, Astle WJ, Megy K, Gräf S, Greene D, Shamardina O, et al. Whole-genome sequencing of patients with rare diseases in a national health system. Nature. 2020;583:96–102.
78. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, Chakravarti A, et al. A global reference for human genetic variation. Nature. 2015;526:68–74.
79. Taliun D, Harris DN, Kessler MD, Carlson J, Szpiech ZA, Torres R, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Nature. 2021;590:290–9.
80. Walter K, Min JL, Huang J, Crooks L, Memari Y, McCarthy S, et al. The UK10K project identifies rare variants in health and disease. Nature. 2015;526:82–90.
81. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536:285–91.
82. Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, et al. REVEL: An Ensemble method for predicting the pathogenicity of rare missense variants. Am J Hum Genet. 2016;99:877–85.
83. Qi H, Zhang H, Zhao Y, Chen C, Long JJ, Chung WK, et al. MVP predicts the pathogenicity of missense variants by deep learning. Nat Commun. 2021;12:510.
84. Kass RE, Raftery AE. Bayes Factors. J Am Stat Assoc. 1995;90:773–95.
85. Yang Y, Muzny DM, Xia F, Niu Z, Person R, Ding Y, et al. Molecular findings among patients referred for clinical whole-exome sequencing. JAMA. 2014. https://doi.org/10.1001/jama.2014.14601.
86. Cipriani V, Pontikos N, Arno G, Sergouniotis PI, Lenassi E, Thawong P, et al. An improved phenotype-driven tool for rare mendelian variant prioritization: benchmarking exomiser on real patient whole-exome data. Genes-basel. 2020;11:460.
87. Sarwal V, Niehus S, Ayyala R, Chang S, Lu A, Darci-Maher N, et al. A comprehensive benchmarking of WGS-based structural variant callers. Biorxiv. 2020:2020.04.16.045120.
88. Zare F, Dow M, Monteleone N, Hosny A, Nabavi S. An evaluation of copy number variation detection tools for cancer using whole exome sequencing data. Bmc Bioinformatics. 2017;18:1–13.
89. Swaminathan GJ, Bragin E, Chatzimichali EA, Corpas M, Bevan AP, Wright CF, et al. DECIPHER: web-based, community resource for clinical interpretation of rare variants in developmental disorders. Hum Mol Genet. 2012;21:R37–44.
90. Yandell MD, Majoros WH. Genomics and natural language processing. Nat Rev Genet. 2002;3:601–10.
91. Kingsmore SF, Cakici JA, Clark MM, Gaughran M, Feddock M, Batalov S, et al. A randomized, controlled trial of the analytic and diagnostic performance of singleton and trio, rapid genome and exome sequencing in ill infants. Am J Hum Genet. 2019;105:719–33.

92. Samocha KE, Robinson EB, Sanders SJ, Stevens C, Sabo A, McGrath LM, et al. A framework for the interpretation of de novo mutation in human disease. Nat Genet. 2014;46:1–8.
93. Liu P, Meng L, Normand EA, Xia F, Song X, Ghazi A, et al. Reanalysis of Clinical Exome Sequencing Data. New Engl J Med. 2019;380:2478–80.
94. Wenger AM, Guturu H, Bernstein JA, Bejerano G. Systematic reanalysis of clinical exome data yields additional diagnoses: implications for providers. Genet Med. 2017;19:209–14.
95. Carapito R, Konantz M, Paillard C, Miao Z, Pichot A, Leduc MS, et al. Mutations in signal recognition particle SRP54 cause syndromic neutropenia with Shwachman-Diamond–like features. J Clin Invest. 2017;127:4090–103.
96. Bellanné-Chantelot C, Schmaltz-Panneau B, Marty C, Fenneteau O, Callebaut I, Clauin S, et al. Mutations in the SRP54 gene cause severe congenital neutropenia as well as Shwachman-Diamond–like syndrome. Blood. 2018; 132:1318–31.
97. SIGNAL RECOGNITION PARTICLE, 54-KD; SRP54Online Mendelian Inheritance in Man® (available at https://omim.org/entry/604857?search=srp54&highlight=srp54).
98. Li Z, Zhang F, Wang Y, Qiu Y, Wu Y, Lu Y, et al. PhenoPro: a novel toolkit for assisting in the diagnosis of Mendelian disease. Bioinformatics. 2019;35: btz100.
99. Deisseroth CA, Birgmeier J, Bodle EE, Kohler JN, Matalon DR, Nazarenko Y, et al. ClinPhen extracts and prioritizes patient phenotypes directly from medical records to expedite genetic disease diagnosis. Genet Med. 2019;21: 1585–93.
100. Baker SW, Murrell JR, Nesbitt AI, Pechter KB, Balciuniene J, Zhao X, et al. Automated clinical exome reanalysis reveals novel diagnoses. J Mol Diagn. 2019;21:38–48.
101. Son JH, Xie G, Yuan C, Ena L, Li Z, Goldstein A, et al. Deep phenotyping on electronic health records facilitates genetic diagnosis by clinical exomes. Am J Hum Genet. 2018;103:58–73.
102. G.-R. D. Collaboration, Center RG, Hout CVV, Tachmazidou I, Backman JD, Hoffman JD, et al. Exome sequencing and characterization of 49,960 individuals in the UK Biobank. Nature. 2020;586:749–56.
103. Carey DJ, Fetterolf SN, Davis FD, Faucett WA, Kirchner HL, Mirshahi U, et al. The Geisinger MyCode community health initiative: an electronic health record–linked biobank for precision medicine research. Genet Med. 2016;18: 906–13.
104. Dimmock DP, Clark MM, Gaughran M, Cakici JA, Caylor SA, Clarke C, et al. An RCT of rapid genomic sequencing among seriously ill infants results in high clinical utility, changes in management, and low perceived harm. Am J Hum Genet. 2020;107:942–52.
105. The Exomiser - a tool to annotate and prioritize exome variants (available at https://github.com/exomiser/Exomiser). Last accessed 22 Aug 2021.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

BMC

UNIVERSITY OF HELSINKI

https://helda.helsinki.fi

Helda

Genomics of asthma, allergy and chronic rhinosinusitis : novel concepts and relevance in airway mucosa

Laulajainen-Hongisto, Anu

John Wiley & Sons Ltd.
2020-10-28

Laulajainen-Hongisto, A, Lyly, A, Hanif, T, Dhaygude, K, Kankainen, M, Renkonen, R, Donner, K, Mattila, P, Jartti, T, Bousquet, J, Kauppi, P & Toppila-Salmi, S 2020, 'Genomics of asthma, allergy and chronic rhinosinusitis : novel concepts and relevance in airway mucosa', Clinical and Translational Allergy, vol. 10, no. 1, 45. https://doi.org/10.1186/s13601-020-00347-6

http://hdl.handle.net/10138/321952
10.1186/s13601-020-00347-6

cc_by
publishedVersion

Downloaded from Helda, University of Helsinki institutional repository.

This is an electronic reprint of the original article.

This reprint may differ from the original in pagination and typographic detail.

Please cite the original version.

Laulajainen-Hongisto *et al. Clin Transl Allergy (2020) 10:45*
https://doi.org/10.1186/s13601-020-00347-6

Clinical and
Translational Allergy

REVIEW

Open Access

Genomics of asthma, allergy and chronic rhinosinusitis: novel concepts and relevance in airway mucosa

Anu Laulajainen-Hongisto[1,2†], Annina Lyly[1,3*†], Tanzeela Hanif[4], Kishor Dhaygude[4], Matti Kankainen[5,6,7], Risto Renkonen[4,5], Kati Donner[6], Pirkko Mattila[4,6], Tuomas Jartti[8], Jean Bousquet[9,10,11], Paula Kauppi[3†] and Sanna Toppila-Salmi[3,4†]

Abstract

Genome wide association studies (GWASs) have revealed several airway disease-associated risk loci. Their role in the onset of asthma, allergic rhinitis (AR) or chronic rhinosinusitis (CRS), however, is not yet fully understood. The aim of this review is to evaluate the airway relevance of loci and genes identified in GWAS studies. GWASs were searched from databases, and a list of loci associating significantly ($p < 10^{-8}$) with asthma, AR and CRS was created. This yielded a total of 267 significantly asthma/AR–associated loci from 31 GWASs. No significant CRS -associated loci were found in this search. A total of 170 protein coding genes were connected to these loci. Of these, 76/170 (44%) showed bronchial epithelial protein expression in stained microscopic figures of Human Protein Atlas (HPA), and 61/170 (36%) had a literature report of having airway epithelial function. Gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) annotation analyses were performed, and 19 functional protein categories were found as significantly ($p < 0.05$) enriched among these genes. These were related to cytokine production, cell activation and adaptive immune response, and all were strongly connected in network analysis. We also identified 15 protein pathways that were significantly ($p < 0.05$) enriched in these genes, related to T-helper cell differentiation, virus infection, JAK-STAT signaling pathway, and asthma. A third of GWAS-level risk loci genes of asthma or AR seemed to have airway epithelial functions according to our database and literature searches. In addition, many of the risk loci genes were immunity related. Some risk loci genes also related to metabolism, neuro-musculoskeletal or other functions. Functions overlapped and formed a strong network in our pathway analyses and are worth future studies of biomarker and therapeutics.

Keywords: Asthma, Allergic rhinitis, Airway epithelium, GWAS, Gene ontology, Pathway

Background

Asthma, allergic rhinitis (AR) and chronic rhinosinusitis (CRS) are multifactorial chronic airway diseases that share some common pathogenetic mechanisms. AR is caused by allergen binding to specific IgE in the nasal mucosa of a sensitized individual, leading to inflammation and symptoms of allergy. The prevalence of AR has increased in the Western countries over the last few decades and it nowadays has been estimated to affect up to 10–25% of the population [1]. Asthma is a chronic pulmonary disease with airway inflammation, bronchial hyperresponsiveness and recurrent, reversible airflow obstruction. Exacerbations are common both in asthma and in CRS, which is a chronic symptomatic inflammation of the sinonasal tract. Asthma and CRS both affect

---

*Correspondence: Annina.lyly@hus.fi
†Anu Laulajainen-Hongisto and Annina Lyly shared first authorship
†Paula Kauppi and Sanna Toppila-Salmi shared last authorship
[1] Department of Otorhinolaryngology–Head and Neck Surgery, University of Helsinki and Helsinki University Hospital, Kasarmikatu 11-13, P.O.Box 263, 00029 HUS Helsinki, Finland
Full list of author information is available at the end of the article

© The Author(s) 2020. This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

about 3–10% of the Western population [2–4]. Risk factors for asthma, AR and CRS include genetic predisposition, other allergic diseases, infections and environmental factors including exposure to tobacco smoke and air pollution [5–7].

Many of the environmental risk factors in the pathogenesis of asthma and CRS are linked to disrupted interplay between epithelial barriers, particles, allergens and microbes [8, 9]. Type 2 biased inflammation with recruitment of eosinophils, basophils, and T-cells, and release of cytokines is common in atopic asthma and AR [10, 11]. Epithelial cells are in contact with microbes, which increasingly have been shown to have a role in inflammatory diseases [12, 13]. Recent studies have also found that altered airway microbiome composition might be associated with asthma [14], seasonal AR [15–17], or children with rhinitis [18].

Genetic inheritance has been estimated to explain 25–80% of asthma risk [19] and up to 90% of AR risk [20]. The genetic predisposition of CRS seems to vary according to CRS type. Although an increased risk is associated with both types, the familial risk of CRS with nasal polyps (CRSwNP) has been found significantly higher than that of CRS without nasal polyps (CRSsNP) in a population based study conducted in Utah [21].

Large collaborative twin studies and GWAS projects have helped establishing genetic components for asthma, CRS and AR. Kim et al. [22] summarized the results of 42 GWASs of asthmatic subjects and controls and asthma-related traits. The most replicated loci with genome-wide significant ($p < 5 \times 10^{-8}$) were the cluster of genes at the 17q12–21, including *ORMDL3* (orosomucoid-like 3), *GSDMB* (gasdermin B), and *GSDMA* (gasdermin A), specific to childhood-onset disease. The next three loci achieving significant p-values included loci 2q12 in the vicinity of several interleukin receptor genes, namely *IL1RL1, IL1RL2*, and *IL18R1*, a region on 5q22 that contains the mitochondrial solute carrier gene *SLC25A46* and the hemopoietic cytokine gene *TSLP* and a complex region located within the major histocompatibility locus 6p21. While these multigene loci are challenging to dissect, it is notable that *IL1RL1* encodes the receptor for IL-33. The gene that encodes IL-33 is separately implicated in the genetic etiology of asthma through the fifth most replicated locus on chromosome 9p24.

Like many other complex diseases, the development of asthma or AR requires genetic predisposition and appropriate timing of environmental exposures. GWASs have been able to identify and replicate several significant risk regions in large sample sets [10]. Among already known important asthma loci, GWASs have also revealed previously undescribed and unexpected genetic components, highlighting the method's freedom of preconceptions [23].

Risk factors and pathogenetic mechanisms of allergic diseases are also interrelated and share partly same mechanisms, which is why we took this broad approach. Our aim was to review GWASs identified asthma, AR and CRS related genes, and to evaluate their relevance in airway mucosal functions. This review is based on an extensive literature search, and several database searches (Fig. 1).

GWAS catalog-search

The GWAS catalog containing 11,598 unique SNPs, was downloaded from the National Human Genome Research Institute (NHGRI) website (https://www.ebi.ac.uk/gwas/) on January 18, 2018. We created a list of 267 SNPs associating significantly (p < 10 exp -8) with asthma and/or AR provided in the Additional file 1: Table S1. There were no SNPs associating significantly with CRS. Gene symbols were mapped onto chromosomes by using Ensemble Karyotype viewer (https://www.ensembl.org/Homo_sapiens/Location/Genome) (Additional file 1: Table S1). Of these 267 SNPs, we selected the SNPs which were assigned to a protein coding gene or those reported to have a protein coding gene as the nearest gene [24]. Using this strategy, we identified a total of 170 protein coding genes (Additional file 1: Table S2). Of them, 21 genes were connected to several SNPs and/or identified in different studies. A Manhattan plot (https://biorender.com/) of SNPs was generated (Fig. 2), showing that susceptibility genes were distributed to all chromosomes (except sex chromosomes). Several genes were found to locate in chromosomes 1, 5, 6, or 17.

Database and literature search of airway expression of the protein expressing genes

Information about the 170 protein coding genes in Gene Cards, NCBI Gene Expression Omnibus (GEOaccession: GSE5057 and GSE40364), and Human Protein Atlas (HPA) (https://www.proteinatlas.org/) was examined (Fig. 3). We also collected lung, bronchial, and nasopharyngeal expressions of these genes from the Genotype tissue expression portal (GTEx), expressed as Reads Per Kilobase Million (RPKM) (Additional file 1: Table S2).

Nasopharyngeal/bronchial protein expression information was obtained from immunohistochemically stained photomicrographs of HPA, and the staining intensity was semiquantitatively scored as 0–3 (0 = no, 1 = mild, 2 = moderate, 3 = strong staining). A total of 76/170 (44%) of these genes showed bronchial epithelial protein expression and 69/170 (41%) showed nasopharyngeal epithelial protein expression in stained microscopic figures

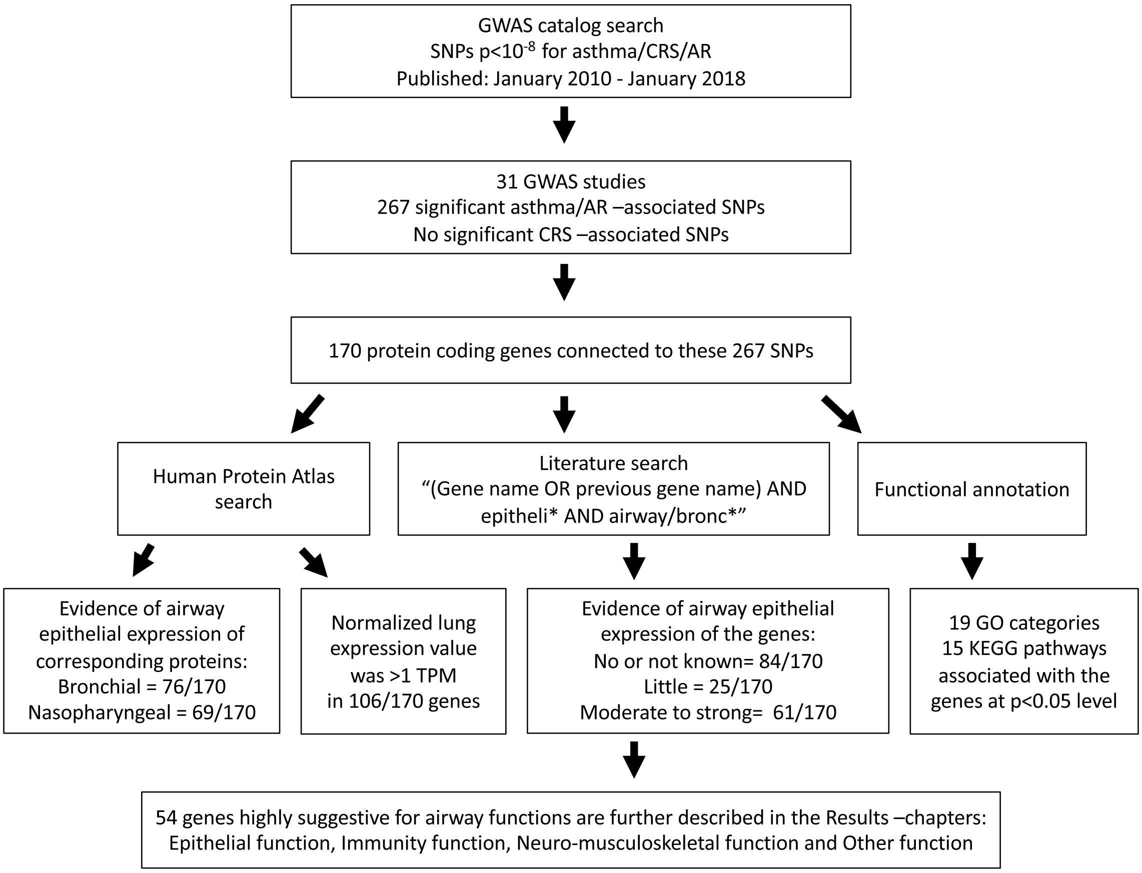

GWAS catalog search
SNPs p<10^{-8} for asthma/CRS/AR
Published: January 2010 - January 2018
31 GWAS studies
267 significant asthma/AR –associated SNPs
No significant CRS –associated SNPs
170 protein coding genes connected to these 267 SNPs
Human Protein Atlas search
Literature search "(Gene name OR previous gene name) AND epitheli* AND airway/bronc*"
Functional annotation
Evidence of airway epithelial expression of corresponding proteins: Bronchial = 76/170 Nasopharyngeal = 69/170
Normalized lung expression value was >1 TPM in 106/170 genes
Evidence of airway epithelial expression of the genes: No or not known= 84/170 Little = 25/170 Moderate to strong= 61/170
19 GO categories 15 KEGG pathways associated with the genes at p<0.05 level
54 genes highly suggestive for airway functions are further described in the Results –chapters: Epithelial function, Immunity function, Neuro-musculoskeletal function and Other function

Fig. 1 Flow chart of the study. GWASs were searched from databases (https://www.ebi.ac.uk/gwas/) and a list and database of SNPs associating significantly ($p < 10^{-8}$) with asthma, AR and CRS was formed in 1/2018. Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) annotation analysis was performed for protein coding genes connected to SNPs. Airway epithelial expression of genes and corresponding proteins were evaluated by using Gene Cards, Human Protein Atlas (HPA), and literature search from PubMed. Other potential airway functions of the selected set of genes were evaluated from database /literature search. Abbreviations: *AR* allergic rhinitis, *CRS* chronic rhinosinusitis, *GO* Gene Ontology, *GWAS* genome-wide association study, *HPA* Human Protein Atlas, *KEGG* Kyoto Encyclopedia of Genes and Genomes, *SNP* single nucleotide polymorphism, *TPM* Transcripts Per Million

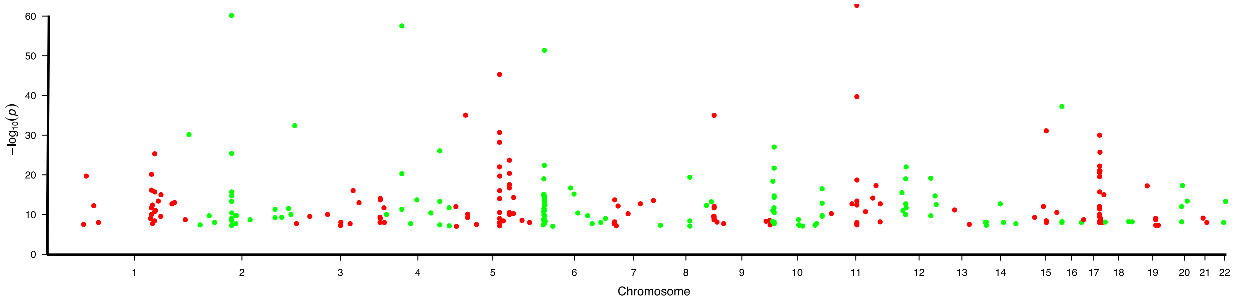


Fig. 2 Manhattan plot (https://biorender.com/) of the SNPs that were significantly associated (p < 5 × 10–8) with asthma/AR in GWASs

(Fig. 3a). The proportion of genes that did not have protein expression figures available was 55% in bronchial and 59% in nasopharyngeal region (Fig. 3a). All available stained microscopic figures showed airway expression, and when scoring semiquantitatively the staining intensity, the proportion of genes showing moderate protein

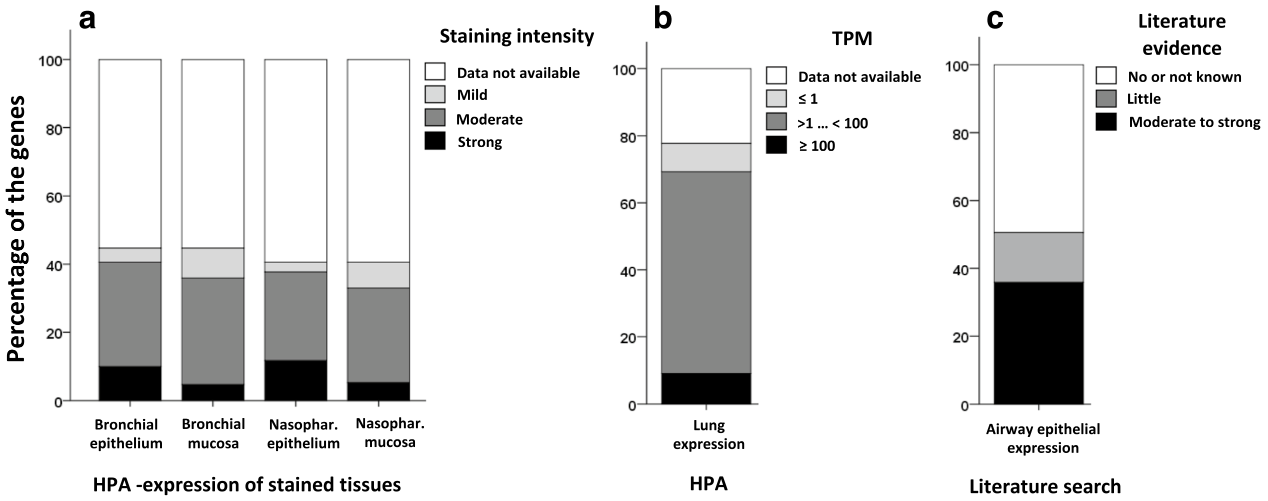


Fig. 3 Database and literature analysis of airway expression of the corresponding proteins of the protein coding GWAS-level genes (n = 170) associating with asthma/AR. **a** Human protein Atlas (HPA) was used to search for photomicrographs of the proteins and their staining intensity was semi-quantitatively evaluated by two observers. **b** HPA results of lung expression in Transcripts per million (TPM). **c** Literature search results of airway epithelial expression of these genes. Pubmed search was performed by using search terms "(Gene name OR alias) AND epitheli* AND airway/bronc*", Nasophar = Nasopharyngeal

expression intensity of all available figures was 68% in bronchial and 64% in nasopharyngeal regions (Fig. 3a). Similarly, the proportion of genes with overall expression was 45% in bronchial and 41% in nasopharyngeal mucosa (Fig. 3a). When evaluating the normalized expression values, e.g. transcripts per million (TPM) in HPA, 51/170 (30%) genes did not show data, 106/170 (62%) genes showed TPM > 1 and 8% genes showed TPM value of < 1 (Fig. 3b). The exact TPM values are shown in the Additional files 1, 2: Tables S1 and S2. When evaluating Genotype-Tissue Expression (GTEx) from HPA, the expression levels for the vast majority of these genes were not available (Additional files 1, 2: Tables S1 and S2).

A systematic literature search in PubMed was performed for the 170 protein coding genes, by using as search terms "Gene name (or any of its aliases) AND epitheli* AND airway/bronc*" and information of airway epithelial expression was scored as 0–2 (0 = no evidence of airway expression, 1 = maybe, 2 = yes/ubiquitous expression) (Additional files 1, 2: Tables S1–S2). A total of 25/170 (15%) genes showed little and 61/170 (36%) showed moderate to strong evidence of airway epithelial protein expression, whereas 49% of genes showed no/not known evidence of airway epithelial protein expression (Fig. 3c).

Functional annotation

We performed functional annotation of the identified genes by R, using a package called clusterProfiler (https://bioconductor.org/packages/release/bioc/html/clusterProfiler.html) (Fig. 4) [25]. The SNPs (Additional file 1: Table S1) that were connected to the protein coding genes were used. If several SNPs were in/near to the same gene, only the first SNP of the list of the Additional file 1: Table S1 was used in functional annotation. Noncoding genes related to these asthma/AR -associated SNPs were excluded in functional annotation. Altogether, there was a total of 155 genes showing functional annotation (Fig. 4). We identified 19 Gene ontology (GO) categories (functional protein categories) that were significantly ($p < 0.05$) enriched among these genes, such as cytokine production, cell activation, leukocyte differentiation, regulation of cell adhesion, leukocyte proliferation, adaptive immune response, antigen receptor signaling pathway and regulation of inflammation response (Fig. 4a). Gene network analysis showed strong interaction between these GO-categories, indicating a strong regulation network between the genes (Fig. 4b). We also identified 15 Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways that were significantly ($p < 0.05$) enriched among the genes, such as Th17 cell differentiation, Th1 and Th2 cell differentiation, Human T-cell leukemia virus 1 infection, Inflammatory bowel disease, Epstein Barr virus infection, Tuberculosis, Hematopoietic cell lineage, JAK-STAT signaling pathway, and asthma (Fig. 4c). JAK-STAT signaling pathway was the most significantly enriched pathway (Fig. 4c).

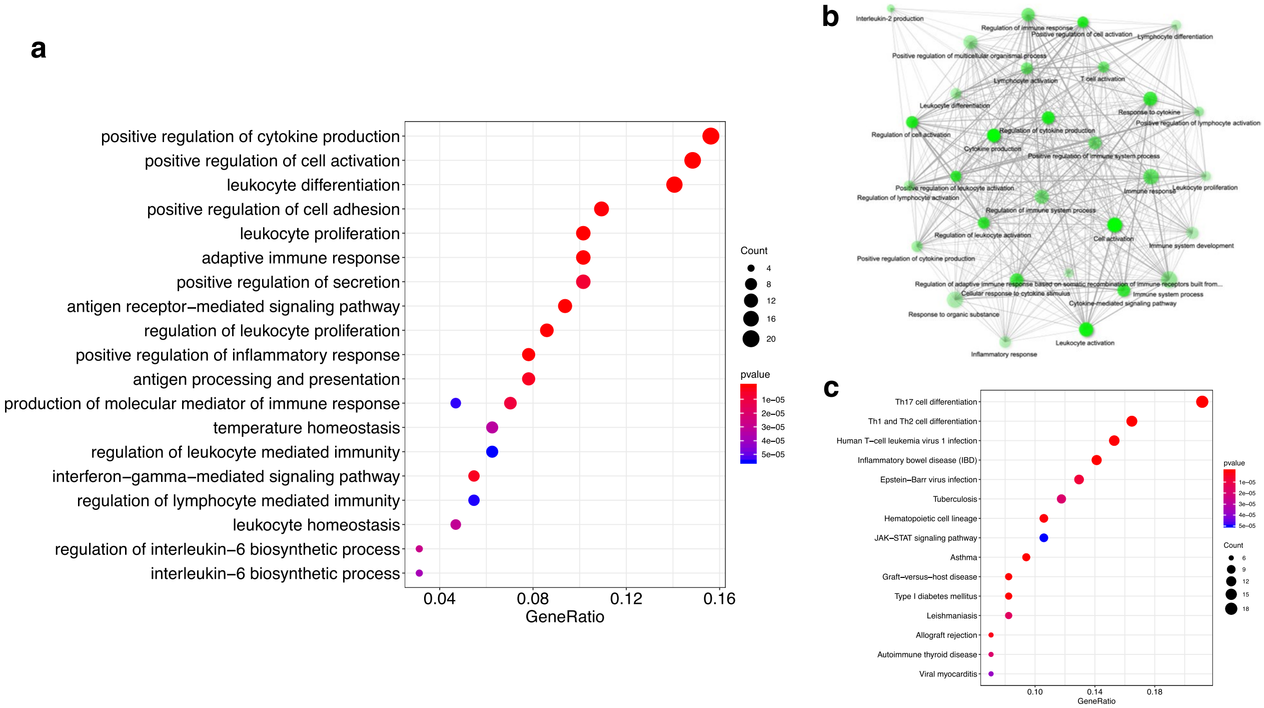


Fig. 4 Functional annotation of the genes corresponding to the SNPs that were significantly associated with asthma/AR in GWASs. We used the whole list of SNPs (Ensemble codes) shown in the Additional file 1: Table S1. Of these, only SNPs that were connected to protein coding genes were used in this analysis, and only one unique gene associated with one or several SNPs were used. Noncoding genes related to these asthma/AR -associated SNPs were excluded in functional annotation. The total number of genes of this functional annotation was 155 genes. **a** Gene ontology (GO) categories (functional protein categories) that were enriched. **b** Gene function network interaction of the GO-categories. The network interaction shows strong interaction between the GO-categories and indicating regulation functions between each other. **c** Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways enriched among 155 genes. Shades of blue and red indicate significance of the enrichment (all were significant at level $p < 0.05$), and the size of the dot represents gene count. X-axis represents the number of genes belonging to the particular category / total number of observed genes ($N = 155$)

Literature review

We finally ranked genes that were highly suggestive of having asthma/AR-relevant functions. The selection was performed based on (a.) the significance (p-value) of particular SNP in GWASs, or (b.) multiple replication of particular SNP or gene in GWASs, or (c.) large literature knowledge of airway function of a particular gene (Additional file 1: Table S1). The selected list of genes is shown in bold text in Table 1. We searched expression knowledge from literature for this smaller set of genes and categorized the genes based on their potential function in the airways in four main groups: genes related to epithelial function, immunity function, neuro-musculoskeletal functions, and other functions. Although some functional groups overlap, genes are only reported in one functional group. In the following text we will discuss the most relevant genes we found in these searches. In Fig. 5 we summarize the main airway functions related to the genes we evaluated most relevant for functions in nasal mucosa during AR and in bronchial mucosa during asthma.

Epithelial function—related genes

CHI3L1 (Chitinase-3-like protein 1) gene encodes chitinase like protein (YKL-40), which is involved in inflammation and tissue remodeling [26]. A study combining GWAS with serum YKL-40 measurement, involving 632 members of the Hutterite population of European decent (living in South Dakota) (age range 6–92 years; asthma in 11.5%, atopy in 41.2%), have shown *CHI3L1* (SNP rs4950928, -131C->G) to be a susceptibility gene for asthma, bronchial hyperresponsiveness, and reduced lung function [27]. Elevated circulating YKL-40 levels have been shown to be a biomarker for asthma [27, 28], they have been shown to correlate with asthma severity,

Table 1 The list of all reported genes associating with asthma/AR/CRS GWAS SNPs at the level of P < 10 exp -8. The reviewed genes are shown in bold

Chr	Genes
1	SFPQ, ZMYM4, **RUNX3**, RERE, TNFRSF14, FAM213B, C1orf54, MRPS21, **FLG,** IL6R, RORC, RPTN, HRNR, **PYHIN1**, DARC, **FCER1A, OR10J3**, NDUFS2, FCER1G, CD247, FASLG, TNFSF18, TNFSF4, CRB1, DENND1B, **CHI3L1**, ITPKB
2	ASB3, SOCS, JUND, CEBPB, IL18R1, **IL1RL1**, IL1RL2, BCL2L11, ANAPC1, IL1B, KYNU, ARHGAP15, **PLCL1**, IKZF2, CCL20, DAW1, INPP5D, D2HGDH
3	RYBP, GLB1, IL5RA, **ABI3BP**, FAM172B, TRMT10C, SLC15A2, GATA2, RASA2, **BCL6, LPP**, DLG1, FBXO45, CEP19
4	**TLR1, TLR6, TLR10,** STX18, MSX1, SRIP1, **GC**, MANBA, **ADAD1, IL2, IL21**, GAB1
5	DAB2, PTGER4, **IL7R**, FBXL7, FAM105A, PDE4D, TMEM232, **SLC25A46, TSLP, WDR36**, CAMK4, TNFAIP8, C5orf56, **IL13, RAD50**, IL5, DIAPH1, NDFIP1, LMAN2, RGS14
6	GRM4, HGMA1, ITPR3, BTNL2, C6orf10, HLA-DPB1, HLA-DOA, HLA-DPA1, **HLA-DQA1, HLA-DQA2, HLA-DQB1**, HLA-DRA, BTNL2, NOTCH4, **PBX2**, HLA-B, MICA, HLA-C, NCR3, AIF1, **PSORS1C1, TNXB**, CREBL1, HLA-A, HLA-G, HLA-J, BACH2, ATG5, PTPRK, TNFAIP3, ARID1B, RNASET2
7	C7orf72, IKZF1, JAZF1, NPY, FERD3L, **ITGB8**, ABCB5, GSAP, **CDHR3**
8	TUSC3, ZBTB10, TPD52, **SLC30A8, MYC**
9	EQTN, TEK, MOB3B, JKAMPP1, TYRP1, JAK2, RANBP6, **IL33**, PHF19, TRAF1, C9orf114, LRRC8A, PTGES
10	**GATA3**, SFTA1P, AKR1E2, IL2RA, ZNF365, JMJD1C, REEP3, PSAP, HPSE2, C10orf95, ACTR1A, TCF7L2
11	DBX1, NAV2, HTATIP2, PRMT3, **AP5B1, OVOL1**, WNT11, **LRRC32, C11orf30**, SESN3, FAM76B, LAYN, SIK2, DDX6, CXCR5, KIRREL3-AS3, ETS1
12	HDAC7, **AQP2**, CDK2, SUOX, IKZF4, **STAT6**, NAB2, ATXN2, SLC22A5, C12orf65, CDK2AP1, SPPL3, HNF1A-AS1
13	FOXO1, PIBF1, KLF5
14	PSMA6, FOXA1, TTC6, RAD51B, JDP2, BATF, RCOR1, TRAF3
15	RTF1, ITPKA, **RORA, SMAD3**, IQGAP1
16	CLEC16A, RMI2, LITAF
17	SMTNL2, ALOX15, GRB7, **GSDMA, GSDMB**, CRKRS, ORMDL3, PERLD1, **IKZF3**, PNMT, **PSMD3**, ZPBP2, **CCR7, SMARCE1**, STAT5B, MAP3K14, ARHGAP27, ZNF652
18	LPIN2, DYNAP, RAB27B, TNFRSF11A
19	SLC7A10, CEBPA, , **ZNF614, ZNF841, ZNF432**, ZNF776
20	NFATC2, **ZNF217**, RTEL1
21	RUNX1, SIK1
22	IL2RB, TEF, TOB2

thickening of subepithelial basement membrane, and inversely with lung function [26].

CDHR3 (Cadherin related family member 3) gene is a transmembrane protein with six extracellular cadherin domains [29]. Cadherins are highly expressed in airway epithelium, and are involved in cell adhesion, epithelial polarity, cell–cell interactions and differentiation [29]. *CDHR3* is expressed in FOXJ1-expressing ciliated cells, which are also the targets of Rhinovirus C (RV-C) binding [30]. GWAS was performed for blood samples of 1173 Danish children (2–6 years) with recurrent acute hospitalizations for asthma, and *CDHR3* (SNP rs6967330; p.Cys529Tyr) has been identified as a susceptibility gene for early childhood asthma with severe exacerbations [29, 31]. Since asthma exacerbations are often caused by infections, it is possible that *CDHR3* variations increase susceptibility to infections, and exacerbations, because of disrupted epithelial integrity [31].

Chromosome 17q21 is an area of interest for asthma and contains a cluster of genes linked to asthma in several GWAS studies, including the *GSDMB* (Gasdermin B) and *IKZF3* (Ikaros family zinc finger protein 3) genes [32]. Chromosome 17q21 has also been linked to inflammatory bowel disease, primary biliary chirrosis, and type 1 diabetes mellitus [32, 33].

Gasdermin A (GSDMA) and B (GDSMB) belong to a family of pore-forming proteins, causing membrane permeabilization and pyroptosis, which is a lytic pro-inflammatory cell death type [34]. Gasdermins are involved in inflammation and cell death, in several herediatry diseases, auto-inflammatory diseases and cancer [34]. *GSDMB* is highly expressed in ciliated airway epithelial cells [35]. Associations have been shown between *GSDMA* gene and asthma [36, 37] and, between *GSDMB* gene and asthma [38] or early childhood asthma with severe exacerbations [39]. Moffat et al. 2010 found an independent association with childhood-onset asthma and *GSDMA* gene (rs3894194, G-> A) in their large consortium-based GWAS of asthma in children and adults from several different populations [36]. Ferreira et al. 2014 found in their GWAS meta-analysis including children and adults from different populations, that *GSDMA* gene (rs7212938, G) was associated with risk of asthma without hay fever (OR 1.14, 95% CI 1.07–1.22) and less

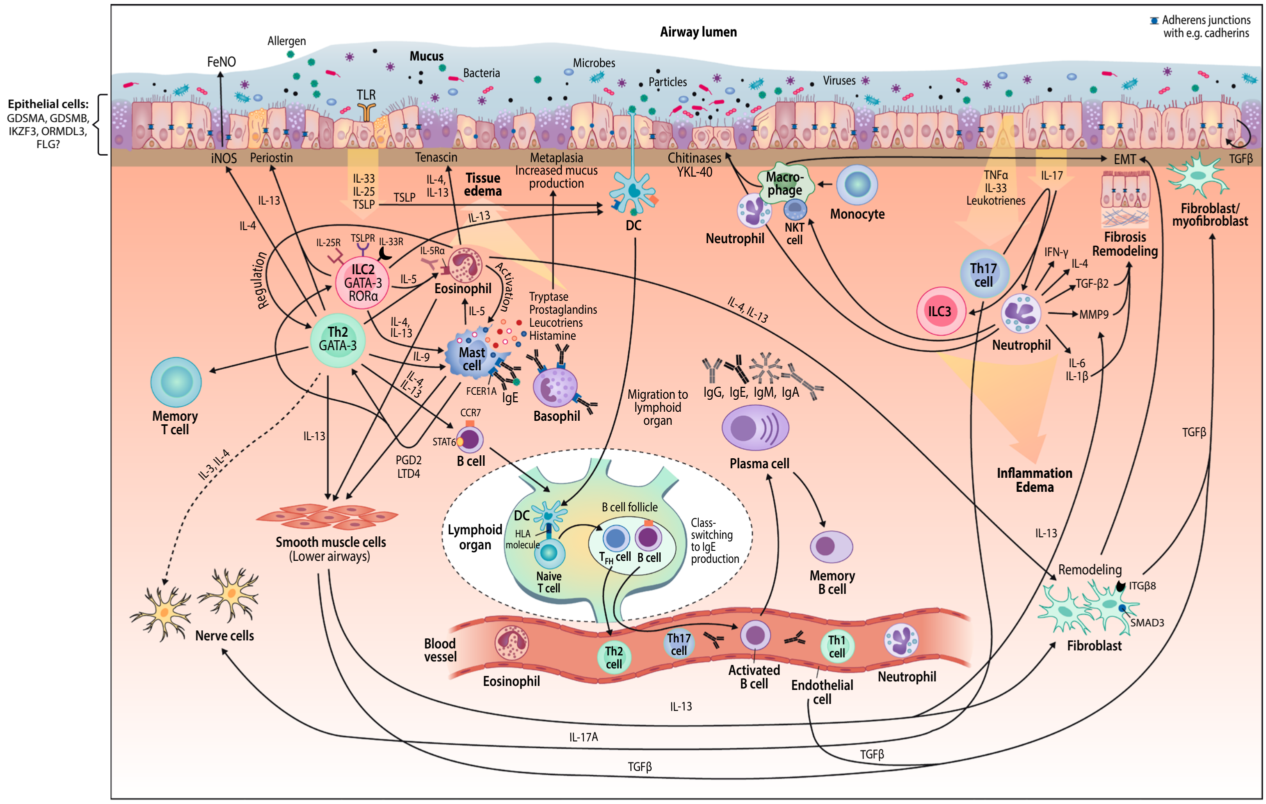

Airway lumen
Adherens junctions with e.g. cadherins
Epithelial cells:
GDSMA, GDSMB, IKZF3, ORMDL3, FLG?
Allergen
FeNO
Mucus
Bacteria
TLR
Microbes
Particles
Viruses
iNOS
Periostin
Tenascin
Metaplasia
Increased mucus production
Chitinases
YKL-40
EMT
TGFβ
IL-33
IL-25
TSLP
TSLP
IL-4,
IL-13
Tissue edema
IL-13
IL-13
IL-4
DC
Macro-phage
Monocyte
Neutrophil
NKT cell
TNFα
IL-33
Leukotrienes
IL-17
Fibrosis
Remodeling
Fibroblast/
myofibroblast
Regulation
IL-25R
TSLPR
IL-33R
ILC2
GATA-3
RORα
IL-5
IL-5Rα
Eosinophil
Activation
IL-5
Tryptase
Prostaglandins
Leucotriens
Histamine
IL-4,
IL-13
Th2
GATA-3
IL-9
Mast cell
FCER1A
IgE
Basophil
IL-4,
IL-13
IL-4, IL-13
Th17 cell
ILC3
Neutrophil
IFN-γ
IL-4
TGF-β2
MMP9
IL-6
IL-1β
Memory
T cell
CCR7
STAT6
B cell
IL-13
PGD2
LTD4
IL-3, IL-4
Migration to lymphoid organ
IgG, IgE, IgM, IgA
Plasma cell
Inflammation
Edema
TGFβ
Smooth muscle cells
(Lower airways)
Lymphoid organ
DC
HLA molecule
Naive T cell
B cell follicle
T_{FH} cell
B cell
Class-switching to IgE production
Memory
B cell
IL-13
Remodeling
ITGβ8
SMAD3
Fibroblast
Nerve cells
Blood vessel
Eosinophil
Th2 cell
Th17 cell
Activated B cell
Th1 cell
Endothelial cell
Neutrophil
IL-13
IL-17A
TGFβ
TGFβ

Fig. 5 Postulated mechanisms of GWAS-related genes in lower airways during asthma. Abbreviations: *CCR7* C-C-motif chemokine receptor 7, *DC* Dendritic cell, *EMT* Epithelial-mesenchymal transition, *FCER1A* Fc fragment of IgE receptor 1A, *FeNO* Fractional exhaled nitric oxide, *FLG* Filaggrin, *GATA-3* GATA binding protein 3, *GDSMA* Gasdermin A, *GDSMB* Gasdermin B, *HLA* Human leukocyte antigen, *IgA* Immunoglobulin A, *IgE* Immunoglobulin E, *IgG* Immunoglobulin G, *IgM* Immunoglobulin M, *IKZF3* Ikaros family zinc finger protein 3, *IL* Interleukin, *ILC2* Group 2 innate lymphoid cells, *ILC3* Group 3 innate lymphoid cells, *INFγ* Interferon gamma, *iNOS* Inducible nitric oxide synthase, *ITGβ8* Integrin Subunit Beta 8, *LTD4* Leukotriene D4, *MMP9* Matrix metalloproteinase 9, *NKT cells* Natural killer T cells, *ORMDL3* Orosomucoid-like 3, *PGD2* Prostaglandin D2, *RORα* Retinoid-Related Orphan Receptor Alpha, *SMAD3* SMAD family member 3 (=Mothers against decapentaplegic homolog 3), *STAT6* Signal transducer and activator of transcription 6, *TFH cell* T follicular helper cell, *TGFβ* Transforming growth factor beta, *Th1* T helper type 1, *Th2* T helper type 2, *Th17* T helper type 17, *TLR* Toll-like receptor, *TNFα* Tumor necrosis factor alpha, *TSLP* Thymic stromal lymphopoietin, *YKL-40* Chitinase like protein, *Remodeling* Smooth muscle and/or fibroblast proliferation, fibrosis, EMT etc.

with risk of hay fever without asthma (OR 1.02, 95% CI 0.98–1.06), suggesting that it is a stronger risk factor for asthma than hay fever [37].

The *GSDMB* gene (rs11078927, C->T) has been linked with asthma by Torgerson et al. 2011, in a GWAS meta-analysis including children and adults from different populations [38]. Also Bønnelykke et al. 2014, linked *GSDMB* gene (rs2305480, G) with early childhood asthma with severe exacerbations in their GWAS of Danish children [39].

A large GWAS of European 180,129 adults/children with asthma and/or AR and/or eczema and 180,709 healthy controls showed that *GSDMB* gene (rs921650 A) is a stronger risk factor for asthma or hay fever than for eczema [40]. A meta-analysis of GWASs of self-reported pollen, dust-mite or cat allergy of 22 012 allergic subjects and 31 850 healthy controls showed that *GSDMB* gene (rs9303280, T->C) was most strongly associated with asthma [41]. A meta-analysis of GWAS of 2144 asthmatic Puerto Ricans and 2893 healthy controls (adults & children) found *GSDMB* gene (rs2305480, G—>A and rs11078927, C—>T) to be associated with asthma [42]. *GSDMB* was highly expressed in nasal epithelial brushings of Puerto Rican children [43], and in primary bronchial epithelium of asthmatic lung [44]. A Dutch GWAS included 920 physician diagnosed asthmatic subjects with bronchial hyperresponsiveness and 980 healthy controls, from northern Netherlands, both children and adults. They compared the GWAS results to prior GWASs, and also performed lung tissue eQTL analysis of the top SNPs replicated in the GWAS analysis [45]. The 17q21 locus achieved genomewide significance, with *GSDMB* (rs8067378, G; rs2305480, A; rs2290400, C; rs7216389, C), also in eQTL analysis, *GSDMB* showed larger effect sizes compared to prior published literature [45]. Exactly how gasdermins affect asthma risk remains unknown, membrane permeabilization and pyroptosis may have an effect in this [35]. When GSDMB protein is cleaved by inflammatory caspase-1 to release its N-terminal fragment, potent pyroptotic cell death had been shown to be induced in airway epithelial cells [35]. There is some evidence that a *GSDMB* splice variant, associated with lower asthma risk, causes an exon deletion leading to GSDMB protein losing its ability to induce pyroptosis in airway epithelial cells, possibly reducing asthma risk [35].

Ikaros family zinc finger protein 3 (IKZF3)s are transcriptional factors involved in lymphocyte differentiation [46] and are expressed in human airway epithelial cells [47]. In a GWAS of Dutch asthmatics with bronchial hyperresponsiveness, *IKZF3* (rs9303277, T) was one of the top findings, and also significant in the lung tissue eQTL analysis [45]. In the meta-analysis of GWAS of asthma in Puerto Ricans (children and adults), the only locus that achieved genome-wide significance for asthma risk was *IKZF3*, (rs907092) at chromosome 17q21 [42]. *IKZF3´*s role in asthma pathogenesis is unknown.

Zinc finger proteins are involved in many cellular processes, and in the development and differentiation of several tissues [48–50]. They are involved in tumorigenesis, cancer progression and metastasis formation e.g. in breast cancer, but can also act as tumor suppressor genes, and are also involved in neurodegeneration, skin diseases (e.g. psoriasis) and diabetes mellitus [48–50]. In the large GWAS of broad allergic disease phenotype (asthma and/or hay fever and/or eczema) by Ferreira et al. [20], *ZNF217* (Zinc Finger Protein 217) gene was one of the identified loci containing genetic risk variants independently associated with the risk of allergic disease [40]. In a GWAS study of Caucasian asthmatic children of which 172 were treated with budesonide and 409 with placebo or nedocromil, the *ZNF432* gene (rs3752120, T—>C) variants were associated with inhaled corticosteroids modulating bronchodilatator response, also *ZNF614* (rs2288884, rs3450) and *ZNF841* (rs12460587, G->T, rs3450) were closely associated [51].

Filaggrin (FLG) is a protein that is critical for keratinization and epithelial barrier homeostasis [52]. Filaggrin gene defects are known to increase the risk of allergic sensitization, atopic eczema and AR [53]. It is the most important genetic risk factor for atopic dermatitis [52, 54]. A GWAS of 1563 European children with physician diagnosed asthma and 4054 controls and a replication analysis showed that, the risk for asthma caused by *FLG* variants (R501X and 2282del4) is limited to asthma cases with co-existing atopic dermatitis [55]. *OVOL1* gene encodes a putative zinc finger containing transcription factor that is highly similar to homologous protein in Drosophila and mouse. *OVOL1* regulates FLG expression in atopic dermatitis subjects [56], and in normal human epidermal keratinocytes [57]. It has been suggested that *FLG* mutations might be involved in barrier dysfunction leading to e.g. asthma. However, in immunohistochemical analysis, filaggrin was not found to be expressed in normal upper airway epithelium in a disagreement with this theory [58].

Toll-like receptors (TLRs) are expressed in nasal epithelium [59], and they have an important innate immunity function recognizing external pathogens and activating immune responses [41]. Nasal epithelial *TLR* gene expression levels were not remarkably altered after nasal birch pollen challenge [59] although a decrease in nasal epithelial TLR1 and TLR6 protein expression was detected in birch pollen allergic adults after challenge [59]. In the meta-analysis of GWASs that have been performed on both children and adult populations with self-reported allergy, the found shared susceptibility loci with asthma included (rs2101521, A->G) chromosome 4p14 near *TLR1*, *TLR6* and *TLR10* [41]. In the meta-analysis of GWASs of children and adults with asthma and hay fever, and controls, *TLR1* (rs4833095, T) was associated with the risk of asthma with co-existing hay fever [37]. The *TLR1* gene was also found to be one of the identified loci containing genetic risk variants independently associated with the risk of allergic disease in the large GWAS of broad allergic disease phenotype (asthma and/or hay fever and/or eczema) by Ferreira et al. [40].

SMAD family member 3 (*SMAD3*) is a transcriptional modulator activated by TGFβ and it may regulate homeostatic and healing pathways to epithelial damage [36]. Mice with SMAD3 deficiency have increased amounts of proinflammatory cytokines in their lungs [60]. The GWAS by Moffat et al. showed an independent association between asthma and the *SMAD3* SNP (rs744910, G->A) [36]. In two GWAS meta-analyses, *SMAD3* SNP (rs17228058, A->G) was a susceptibility locus of asthma and self-reported allergy [41] and, asthma with co-existing hay fever, but not asthma alone [37].

The role of epithelial to mesenchymal transition (EMT) has a critical role in airway remodeling. Human eosinophils co-cultured with bronchial epithelial cells induced EMT, suggestive of their role in airway remodeling, with increased expression of TGFβ1 and *SMAD3* phosphorylation in the bronchial epithelial cells [61].

ITGB8 gene encodes Integrin Subunit Beta 8. This protein noncovalently binds to an alpha subunit to form a heterodimeric integrin complex. In general, integrin complexes mediate cell–cell and cell-extracellular matrix interactions and this complex plays a role in human airway epithelial proliferation. High expression levels of

ITGB8 have been associated with high angiogenic and poorly invasive glioblastoma tumors. Inactivation of *ITGB8* in the murine airway has been associated with a reduction in IL-1β–induced airway inflammation and fibrosis, which is due to decreased TGF-β activation [62].

Immunity function–related genes

Immunity related SNPs in asthma are mostly in genes linked to HLA region and type 2 inflammation. The region 6p21 (*HLA* region) is one of the most replicated asthma loci [22]. Several significant SNPs have been associated with class II major histocompatibility antigen (*HLA-DR*) genes including *HLA-DQA1*, *HLA-DQA2* and *HLA-DQB1* [38]. They play a central role in the immune system by presenting peptides derived from extracellular proteins. Class II molecules are expressed in antigen presenting cells, ie. B lymphocytes, dendritic cells and macrophages and are extensively studied because of the association with several autoimmune, infectious and inflammatory diseases [63].

Group-specific Component (*GC*) gene [also known as Vitamin D-binding protein (*VDBP*) gene] on chromosome 4q13 and has been found to associate with asthma in children [64]. The rs7041 G-allele was found with increased risk [OR 2.15, CI 95% (1.32–3.50; $P = 0.002$)] of asthma in codominant, dominant, recessive and allelic models [64]. VDBP carries circulating vitamin D to the target organs, it is a chemotactic factor for leukocytes and macrophage activation, and also has a role in osteoclast activation [64, 65].

Approximately half of the patients with asthma, regardless of the severity of the disease, exhibit type 2 endotype. The endotype is characterized by a predominant activation of Th2 cells that produce cytokines such as interleukins 4, 5, and 13. These interleukins are responsible of Th2 cell differentiation, maturation and release of eosinophils and proliferation of IgE-producing B-cells, respectively [66]. IL-25, IL-33 and TSLP are thought to be master regulators of type 2 inflammation in diseases and they can all activate innate and adaptive immune cells to secrete IL-5 and IL-13 [67].

Interleukin 1 receptor like 1 (*IL1RL1, ST2*) is an important asthma gene and part of a cytokine receptor gene cluster. GWASs have reproducibly found the *IL1RL1* gene to be associated with asthma susceptibility [22]. *IL1RL1* encodes different isoforms of the receptor: IL1RL1-a is a soluble form and IL1RL1-b is a transmembrane receptor. IL1RL1-a functions as a decoy receptor to dampen IL-33 induced signaling. IL1RL1-b together with IL1RAcP forms a heterodimeric transmembrane receptor for its ligand, IL-33. Binding of IL-33 initiates an MyD-88-mediated signaling cascade, releasing pro-inflammatory cytokines IL-4, IL-5 and IL-13 [68]. *IL-33-ST2* (*IL-1RL1*) axis has been regarded as one of the key players also in allergic diseases, asthma and atopic dermatitis [69]. A number of studies have indicated that IL-33 induces the activation and expansion of group 2 innate lymphoid cells (ILC2s) which cause allergic inflammation by producing large amounts of IL-5 and IL-13 [70]. As IL-5 is the main component of eosinophil activation and survival, anti-IL-5 treatments are used to inhibit eosinophilic inflammation. At the moment, three biologics targeting IL-5 signaling are available: mepolizumab and reslizumab, which bind to IL-5 directly reducing the production and survival of eosinophils, and benralizumab, which targets the IL-5 receptor expressed on eosinophils causing a direct destruction of the cell type [71]. Many functional studies of asthma have focused on peripheral blood mononuclear cells, yet there are also some studies on granulocyte functions: a study has showed that persistent high blood neutrophilia was associated with poor asthma control [72].

Region 9p24 also belongs to one of the most replicated asthma loci, associating with the *IL-33* gene. Being one of the major upstream regulators of type 2 inflammation, *IL-33* has been linked to both asthma and allergic inflammation. It also functions as an”alarmin” and is secreted following tissue damage caused for example by an infection [68]. *IL-33* expression in the lungs is increased in asthma [73]. A recent publication describes a rare loss-of-function mutation in *IL-33,* protecting from asthma [74]. Recently, a biologic recombinant protein called IL-33trap was shown to neutralize IL-33 and inhibit acute allergic airway inflammation in a mouse model [75]. Clinical trials are ongoing with several monoclonal antibodies targeting *IL-33/ST2* signaling [76].

IL-13 is a cytokine secreted by activated Th2 cells, and acts as an important mediator of allergic inflammation pathogenesis. It shares a common receptor subunit with IL-4, namely IL-4Ralpha, therefore sharing also many functions with it, including promoting B-cell proliferation and class switch to IgG4 and IgE [77]. Dupilumab, a biologic targeting this common receptor subunit, has been shown to be effective in many allergic diseases including asthma, atopic dermatitis [78] and CRSwNP [79]. Some functions of IL-13 are independent of IL-4, and especially mucus hypersecretion, subepithelial fibrosis and stimulation of matrix metalloproteinases resulting in emphysematous changes in mouse model has been shown to result from the functions of IL-13 [80, 81]. A recent study investigated the association of 236 candidate gene polymorphisms and asthma disease severity, and found only one marker, the rs848 in the *IL-13* gene region, significantly associating with symptom severity in adults with asthma [82].

TSLP (thymic stromal lymphopoietin) gene encodes a cytokine that is expressed mainly in epithelial cells and plays a key role in allergic inflammatory responses [83]. In humans, dendritic cells are the major target for TSLP, having an integral role on promoting Th2 cell responses. TLSP is produced by the airway epithelium in response to inhaled allergens and proinflammatory stressors and has an upstream role in the asthma cascade [83]. Tezepelumab binds to TSLP, inhibiting its stimulating activity on dendritic cells and innate lymphoid cells thus preventing the induction of type 2 cytokines IL-4, IL-5 and IL-13. Anti-TSLP treatment with tezepelumab decreased asthma exacerbations significantly [84] and phase 3 trials are ongoing.

The expression and production of Th2 cytokines IL-4, -5 and -13, have been shown in isolated cell systems and invertebrates to be controlled by the zinc finger transcription factor *GATA3*, which is essential for Th2-cell differentiation and activation and is considered to be the master transcription factor of the Th2 pathway of immune activation [85]. GATA3$^+$ Th2 cells have been observed in specimens from bronchoalveolar lavage and lung biopsies obtained from patients with severe asthma, even after continuous per oral corticosteroid [86]. CD2-Gata3 transgenic mice developed allergic airway inflammation and showed enhanced levels of IL-5 and IL-13 in bronchoalveolar lavage and lung tissue after allergen induction [87]. A novel therapy for the treatment of Th2-driven asthma targeted GATA3, a transcription factor that plays a key role in Th2 cell differentiation, through an inhaled DNA enzyme (DNAzyme) that specifically cleaves and inactivates *GATA3* mRNA. In a study of allergic asthmatic patients with sputum eosinophilia and biphasic early and late asthmatic responses after allergen provocation, inhalation of GATA3-specific DNAzyme once daily for 28 days attenuated both early and late asthmatic responses to allergen provocation when compared with placebo [88].

The protein encoded by *STAT6* gene is a member of the signal transducer and activator of transcription (STAT) family of transcription factors. In response to cytokines IL-4 and IL-13, STAT6 is phosphorylated by the receptor associated kinases, and then form homo- or heterodimers that translocate to the cell nucleus where they act as transcription activators for a large number of genes involved in macrophage polarization [89]. STAT6 has been demonstrated to regulate many characteristic features of lung inflammation common in asthma, including airway eosinophilia, epithelial mucus production, Th2 cell differentiation, and IgE production from B cells [90].

On chromosome 17q21, two intergenic variant SNPs between *CCR7* and *SMARCE1* associate with high p-value in GWAS reported by Ferreira et al. [40]. As mentioned above, the locus 17q21 is the most replicated in asthma GWASs, including many other important asthma genes such as *ORDLM3* and *GSDMB* [22]. The c–c-motif chemokine receptor CCR7 is a member of the G protein-coupled receptor family. It is responsible for the proper recruitment of lymphocytes and mature dendritic cells to lymphoid tissues. Dendritic cells, T-lymphocytes and B-lymphocytes express CCR7 on their surface, and it has been shown to promote the internalization of antigens by DCs, and to regulate cell survival, migration, and to induce dendritic cell maturation [91, 92]. Recent study investigated the effects of *CCR7* knockdown and overexpression on dendritic cell-mediated immune tolerance in the lungs of rats with allergic asthma and found that CCR7 expression levels affected the expression of various cytokines such as IL-12, IL-4, IFN-γ and IgE, as well as the amount of immune cells in the lungs [93].

FCER1A (Fc fragment of IgE receptor 1A), an initiator of the allergic response, is located on chromosome 1q23, next to *OR10J3* coding for an olfactory receptor, and showing more evidence on association. Variants in *FCER1A* has been reported to associate with total IgE levels, allergic sensitization [94] and atopy [95]. Association of *FCER1A* polymorphism with CRSwNP has been studied in North Indian population-based case–control study [96]. Although no significant association was found with CRSwNP alone, a significant association ($P < 0.05$) of rs2427827 SNP with high IgE level CRSwNP patients was revealed [96]. FCER1A has been shown to be expressed in mast cells and basophils as well as in monocytes and dendritic cells, and it has been suggested to have a dual role in IgE-signaling – studies conducted using transgenic mouse models have shown that on one hand, FCER1A expression induces type 2 inflammation in the lungs after viral infection, on the other hand it has been linked to regulatory role in asthma, promoting immune homeostasis (reviewed in [97]).

ADAD1—IL-2/IL-21: Hinds et al. report a SNP rs4145717-T in the 4q27 region that falls in the *ADAD1* gene, but the nearby *IL-2* and *IL-21* genes show more evidence on association [41]. In this region is another SNP associated with allergic rhinitis [98]. The IL-2 and IL-21 cytokines are involved in the regulation of multiple helper T cell types: IL-21 is needed for germinal center formation by generation of T follicular helper cells [99] and IL-2 is required for Th1, Th2 and Th17 cell differentiation [100].

Neuro-musculoskeletal function–related genes

C11orf30-LRRC32 region has been associated with asthma in previous studies [22]. *LRRC32* (also known as *GARP*, glycoprotein A repetitions predominant) (11q13) is expressed especially in lung and placenta and considered to be involved in several processes. It is a surface molecule of T regulatory cells [101] and earlier found in association to atopic dermatitis and induction of tolerance. C11orf30/EMSY is a transcriptional factor associating with tumor suppressor BRCA2, and it has been linked to above-mentioned TSLP activation in eosinophilic esophagitis [102]. In a multicenter population-based study, *C11orf30*-rs2155219 was reported to double the risk of polysensitisation [103]. Polysensitisation is associated to asthma [104]. *C11orf30/EMSY* has also been found a risk locus for both peanut and food allergy [105].

MYC encodes a transcription factor and located in the chromosome 8q24. MYC is a proto-oncogene and involved in Burkitt lymphoma and multiple myeloma and serves as a prognostic factor in acute myeloid leukemia [106]. A study showed that asthmatics have increased MYC expression in peripheral blood ILC [107]. They deleted *c-Myc* from murine lung ILC2 or an ILC2 cell line by CRISPR knockout, and showed reduced proliferation, decreased cytokine production, and reduced expression of many lymphocyte activation genes. In murine model of airway epithelial injury, *Myc* regulated proliferation and Fibroblast growth factor expression in airway smooth muscle [108].

Chromosomal area 6p21 (within the major histocompatibility complex gene) also includes Tenascin XB gene (*TNXB)* which is a member of tenascin family and extracellular matrix glycoproteins, and has musculoskeletal functions [55]. Tenascins are considered to be anti-adhesive and associated to wound healing, they have also been associated to Ehlers-Danlos syndrome and to malignant mesothelioma and are considered to be involved in airway remodeling in asthma [109]. Also the pre-B-cell leukemia homeobox 2 (*PBX2*) gene (rs204993), located in 6p21, is associated to both asthma and AR in Chinese population [110]. *PBX2* is expressed particularly in epithelium [110].

PSMD3 (chromosome 17q21) is a multicatalytic proteinase complex for proteasome and are distributed in many cells and have been reported to associate to atopic march and to atopic eczema and wheeze [111].

RAD50 (chromosome 5q31) is expressed in many tissues and needed for DNA double-strand break repair and other activities essential for cell growth. Further, it has been reported to associate with atopic dermatitis in Korean population [112]. *RAD50* has been shown to be expressed both in bronchial epithelial cells and bronchial alveolar lavage. However, *IL13*, locating in the same chromosomal region, showed more evidence on association with asthma [113].

RORA (chromosome 15q) is expressed especially in skin and adrenal gland and to less extent other tissues and is a member of NR1 subfamily of nuclear hormone receptors interacting in organogenesis and circadian rhythm [114]. Seven *RORA* SNPs were associated with childhood asthma in European populations, and *RORA* show epistasis with *NPSR1* [115]. The group showed in cell models that stimulation of NPSR1 activated RORA-relevant pathway [115], and that NPS induced *RORA* mRNA expression in neuroblast cell line [116].

The *RUNX* gene family encodes Runt-related transcription factors RUNX 1, 2 and 3. *RUNX3* (1p36) is expressed especially in bone marrow, lymph nodes and spleen and is involved in activating or suppressing transcription and is associated to tumor suppressor. *RUNX3* has been found to be hypomethylated and with increased association to inner city asthma in children [117]. RUNX1 is associated with aberrant B cell maturation and is related to acute myeloid or chronic leukemia [118]. Our study group showed that association between maternal smoking exposure and incident asthma in adult offspring was accentuated in offspring who had haplotype rs11702779-AA of *RUNX1* gene [119].

WDR36 (chromosome 5q22) encodes a member of the WD repeat protein family, involved in many cell functions e.g. signal transduction and apoptosis, and it has been associated with asthma and allergy [120–122]. In recent candidate gene analysis in Han Chinese population, it has been shown to associate with AR [123]. In vitro- studies show the involvement of WDR36 in G_q-coupled muscarine, bradykinin and histamine receptor signaling [124], all of which are important modulators of allergic reactions and asthmatic bronchoconstriction.

Other function–related genes

The Solute Carrier family members belong to the mitochondrial transporter family, which have an important role in metabolism [125]. The tissue distribution and cellular/subcellular expression of *SLC25A46* (Solute Carrier family 25 member 46) is ubiquitous [125]. A GWAS meta-analysis of 3933 European adults with allergic rhinitis and 8965 controls, and 2315 subjects with grass sensitization and 10 032 controls showed that a variant (rs2155219) located close to *SLC25A46* was associated with AR/grass sensitization [98]. The large GWAS of broad allergic disease phenotype (asthma and/or hay fever and/or eczema) by Ferreira et al. [40] identified also *SLC25A46* as one of the identified loci containing genetic risk variants independently associated with the risk of allergic diseas. A GWAS including Japanese and Korean pediatric asthma patients and controls, identified *SLC30A8* (Solute Carrier Family 30 member 8) SNP (rs3019885—> T/G) to associate strongly with pediatric asthma. *SLC30A8* is also known as Zinc Transporter 8, it is a zinc efflux transporter, highly expressed only in the pancreas, its variants are associated with diabetes mellitus type 2 [126]. It is not known how the Solute Carrier family members affect the risk of asthma.

LPP (Lipoma-preferred partner) is a member of a protein family regulating cytoskeletal organization, cell motility and mechanosensing and functions as a mediator of transforming growth factor β (TGFβ) induced cell migration and invasion in breast cancer cells [127]. A study group has found that the *LPP* gene (rs9860547, G->A) is a shared susceptibility locus of asthma and self-reported allergy, with the risk allele being protective against allergy in the meta-analysis of GWASs, in children and adults with self-reported allergy and controls [41]. The study group also performed eQTL analysis, the results of which suggested that the *LPP* gene association may be mediated by an effect on *BCL6* (B cell lymphoma 6) expression affecting STAT6-mediated responses on IL-4, IL-13, and IgE class switching [41].

Elevated sputum *PSORS1C1* (Psoriasis Susceptibility 1 Candidate 1) levels have been shown in chronic obstructive pulmonary disease (COPD) [128]. Pyrin and HIN domain family member 1 (*PYHIN1*) is related to initiation of innate immune response via detection of foreign DNA and, PYHIN1 was shown to positively regulate LPS-induced IFN-β and NO production through up-regulating the MyD88-independent signaling pathway in murine macrophage cell model [129]. Aquaporins (AQPs) mediate fast transmembrane transport of water thus regulating fluid balance in the organs [130]. *AQP2* gene encodes kidney´s water channel protein and it mediates urine water concentration and regulates water balance [130]. A mouse model showed that a Chinese herb, Platycodon root, by diffusing the lung can ameliorate the respiratory-function and pathologic changes in the lung tissues, but also regulate urinary output and renal expression of AQP1 and AQP2 [130].

PLCL1 (Phospholipase C-Like 1) is involved in an inositol phospholipid-based intracellular signaling cascade [41]. It has been reported to be related to Circadian entrainment [131] and Crohn's disease [41]. ABI3BP (ABI Family Member 3 Binding Protein) is a extracellular matrix protein and is expressed in multiple organs, including the heart, kidney, lung, pancreas, and placenta, with low-level or variable expression in the spleen, liver, brain, bone, and skeletal muscle [132]. *ABI3BP* gene has been shown to contribute emphysema phenotype in a mouse model that were exposed to cigarette smoke [133]. On the other hand a study showed that knockout of *Abi3bp* in mice does not affect their olfactory function, mental state and NNK-induced lung tumorigenesis [134]. *AP5B1* (Adaptor Related Protein Complex 5 Subunit Beta 1) is associated with Hereditary Spastic Paraplegia [135].

Conclusions

Development of post-GWAS methods are important for characterizing the function of trait-associated loci [136]. Strategies integrating various biological data sets with GWAS results will provide insights into the mechanistic role of associated loci. For example, an integrated GWAS and expression study on AR highlights mitochondrial pathways as a target for further investigation of AR mechanism and treatment [137].

We identified a total of 267 significantly asthma or AR –associating loci from 31 GWAS studies and 170 protein coding GWAS-level risk genes of asthma or AR. Of these about a third had airway epithelial functions in database and literature search. In addition, many genes have been related to immunity functions and in part to neuro-musculoskeletal and other functions in literature. These functions overlapped and also formed a strong network in pathway analyses. Still it is noteworthy that not all SNPs would be asthma markers themselves, or that each locus may lead to pathogenesis of AR or asthma. In addition, in about half of the protein coding genes the expression figures of databases was not yet available. Thus, further functional experiments would be needed to study their putative role in airways.

There is still scarce GWAS-level knowledge of CRS phenotype. After our GWAS catalogue search, a GWAS publication of Islandic and English CRS patients and controls showed that a missense variant in *ALOX15* causing p.Thr560Met alteration in arachidonate 15-lipoxygenase (*15-LO*) confers large genome-wide significant protection against CRSwNP and CRS. p.Thr560Met, carried by 1 in 20 Europeans, was previously shown to cause near total loss of 15-LO enzymatic activity [138]. The authors suggest that the protective effect of this variant is explained by inactivation of 15-LO´s catalytic activity, which leads to reduced production of pro-inflammatory mediators in eosinophils and nasal epithelium [138].

Since asthma and allergies are multifactorial disorders affected by both genetic and environmental factors and with multiple phenotypes, it seems likely that several genes are involved, each with a minor effect. It is expected that different genetic pathways are involved with varying proportions in different populations. Allergic diseases are also heterogeneous. Clinically, some patients have allergic rhinitis (AR) alone whereas others have AR and asthma (with or without other allergic manifestations). Few patients have asthma alone. Using transcriptomics analyses in European MeDALL birth cohorts and RNA sequencing in Puerto Rican children, AR as a single disease was found to be associated with Toll-like receptor gene expression (*TLR*), whereas rhinitis associated with asthma was linked with IL-5 and IL-33, confirming that the two diseases are different [139].

We show that many risk genes relate to leukocyte immunity or epithelial cell functions. Approximately half the patients with asthma, regardless of the severity of the disease, and majority of AR cases, exhibit Th2 endotype [67]. This endotype is characterized by a predominant activation of Th2 cells that produce cytokines such as IL-4, -5, and -13 [67]. There is increasing evidence of airway barrier´s role in airway diseases, such as AR [17]. Although expression profiles in upper and lower airways might differ [140], investigations on nasal epithelial interactions might provide additional knowledge for the lower airway inflammation [43, 141].

Therapeutic implications and future prospects

Since many of the identified risk genes for inflammatory airway diseases are airway epithelial or immunity function related, these functions should have an important role in the search for gene-environmental interactions, biomarkers and future therapeutics. Metabolism and neuro-musculoskeletal related functions also seem to have a central role in the development of asthma and AR.

With knowledge of important cascades in asthma and AR pathogenesis, as revealed also by these GWASs, it has been possible to develop medications that specifically target the key players of these cascades. In the 2010′s, biologics suppressing type 2 inflammation have become an important tool when tackling severe eosinophilic asthma, and recently also severe CRSwNP, and new substances are also under investigation [142].

Asthma endotypes can broadly be divided to type 2 high or type 2 low asthma [143]. Type 2 endotype is defined by the presence of Th2- and ILC2- inflammatory markers and eosinophilia. Type 2 low endotype is not as well defined; it is characterized by the absence of type 2 markers, and by activation of neutrophils, Th1 and/or Th17 cells [143]. Some patients, often with severe asthma, have a mixed population of airway granulocytes (eosinophils and neutrophils), and combined type 2 high and low cytokine signatures, such as IL-17 or IFN-γ [144]. This heterogeneity could explain why fewer genome-scale asthma loci have been identified in type 2 low asthma than in type 2 high asthma [143]. Lack of data also explains why molecules/pathways relevant to type 2 low asthma were discussed less in this review. Nevertheless, there are several mucosal molecules/pathways that are currently under investigation as potential therapeutic targets for type 2 low or mixed type of asthma, these include IL-6 [145] and TLR-3,4,7 [146].

After this study was performed, new GWASs and functional annotation analysis have been published with new interesting asthma loci such as TNF receptor superfamily member 8 (*TNFRSF8*) [143], and Collagen Type XVI Alpha 1 Chain (*COL16A1*) [147], and more studies are to come. This is indeed very important since further functional activity experiments of candidate genes are still needed to understand the dynamic molecular events behind the pathogenesis of airway diseases. In addition, there is a high need in the future to perform GWASs also on CRS, type 2 low asthma, chronic obstructive lung disease, and rare severe airway diseases, in order to broaden our understanding of all inflammatory airway disease subtypes and to discover new pathways/molecules. This knowledge would potentially improve future preventive and therapeutic strategies for inflammatory airway diseases.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s13601-020-00347-6.

Additional file 1: Table S1. The full Excel file (as a supple file) of the results of ours including also the SNPs with no relevant signs of airway epithelial expression, and the PMIDs. There are several rows and they are based on 1) associated transcript for particular gene 2) allele changes 3) Multiple GWAS-studies that have shown the same loci. These explain the over 60 000 rows of the dataset. Abbreviations: N/A = not available, Protein_Atlas_Lung_Expression(TPM02) = 0 = TPM less than 1; 1 = 1 < TPM < 100; 2 = TPM 100 or over, Protein_Atlas_Lung_Expression(TPM) = total TPM value, HPA BrProt exp epi = Bronchial epithelial protein expression of immunohistochemical staining figures of The Human Protein Atlas, HPA Br Prot exp = Bronchial protein expression of immunohistochemical staining figures of The Human Protein Atlas, HPA Nasoph Prot exp epi = Nasopharyngeal epithelial protein expression of immunohistochemical staining figures of The Human Protein Atlas, HPA Nasoph Prot exp = Nasopharyngeal protein expression of immunohistochemical staining figures of The Human Protein Atlas, HPA Nasoph GTEx (RPKM) = The Genotype-Tissue Expression (GTEx) of Nasopharyngeal mucosa of The Human Protein Atlas, RPKM = Reads Per Kilobase of transcript, per Million mapped reads, HPA Bronch GTEx (RPKM) = The Genotype-Tissue Expression (GTEx) of Bronchial mucosa of The Human Protein Atlas, GXA = Gene Expression Database.

Additional file 2: Table S2. Integrated data from GWAS catalog on asthma, allergy/AR and CRS -associating SNPs, the corresponding proteins of which have evidence of human airway epithelial expression from manual literature search. The list of 170 asthma/AR –relevant coding gene names were identified based on GWAS search shown in Fig. 1. We detected that 145/170 genes have reported epithelial gene and/or protein expression based on systematic Human Protein Atlas (HPA) search. We detected that 86/170 genes have at least slight evidence of airway epithelial expression based on literature search, the citations are shown in the list. Abbreviations: Protein_Atlas_Lung_Expression(TPM02) = 0 = TPM less than 1; 1 = 1 < TPM < 100; 2 = TPM 100 or over, Protein_Atlas_Lung_Expression(TPM) = total TPM value, HPA BrProt exp epi = Bronchial epithelial protein expression of immunohistochemical staining figures of The Human Protein Atlas, HPA Br Prot exp = Bronchial protein expression of immunohistochemical staining figures of The Human Protein Atlas, HPA Nasoph Prot exp epi = Nasopharyngeal epithelial protein expression of immunohistochemical staining figures of The Human Protein Atlas, HPA Nasoph Prot exp = Nasopharyngeal protein expression of immunohistochemical staining figures of The Human Protein Atlas, HPA Nasoph GTEx (RPKM) = The Genotype-Tissue Expression (GTEx) of Nasopharyngeal mucosa of The Human Protein Atlas, RPKM = Reads Per Kilobase of transcript, per Million mapped reads, HPA Bronch GTEx (RPKM) = The Genotype-Tissue Expression (GTEx) of Bronchial mucosa of The Human Protein Atlas, GXA = Gene Expression Database, Literat_Resp_epith = Literature search result of Airway expression: 0 = no/not known; 1 = some evidence; 2 = evidence exists, Literat Ref (PMID) = PubMed ID for the Literature search result.

Abbreviations

AR: Allergic rhinitis; CRS: Chronic rhinosinusitis; GWAS: Genome wide association study; GO: Gene Ontology; SNP: Single-nucleotide polymorphism; KEGG: Kyoto Encyclopedia of Genes and Genomes.

Acknowledgements

The authors wish to thank Helena Schmidt for the medical illustration.

Author contributions

ST-S provided the study plan with RR, PM, KD and MK. The data was collected and analyzed by ST-S, TH, KD, MK. ST-S, AL, ALH and PK performed the literature review and wrote the manuscript. All authors critically reviewed the data analyses and the manuscript. All authors read and approved the final manuscript.

Funding

The study was supported in part by research grants from Finnish Medical Foundation, the Finnish Society of Allergology and Immunology, the Jane and Aatos Erkko Foundation, the Finnish Cultural Foundation, Hospital District of Helsinki and Uusimaa (TYH2018103, TYH2019322), Paulo Foundation, the Tampere Tuberculosis Foundation, the Väinö and Laina Kivi Foundation, the Finnish ORL-HNS Foundation.

Availability of data and materials

All data generated or analyzed during this study are included in this published article and its Additional files 1, 2.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

STS has acted as paid consultant for Mylan Laboratories Ltd., Biomedical systems Ltd., Roche Products Ltd., and Sanofi S.A. All other authors declare no conflicts of interest.

Author details

[1] Department of Otorhinolaryngology–Head and Neck Surgery, University of Helsinki and Helsinki University Hospital, Kasarmikatu 11-13, P.O.Box 263, 00029 HUS Helsinki, Finland. [2] Laboratory of Cellular and Molecular Immunology, Institute of Microbiology of the Czech Academy of Sciences, Prague, Czech Republic. [3] Skin and Allergy Hospital, University of Helsinki and Helsinki University Hospital, Helsinki, Finland. [4] Haartman Institute, University of Helsinki, Helsinki, Finland. [5] HUS Diagnostic Center, Helsinki University Hospital, Helsinki, Finland. [6] Hematology Research Unit Helsinki, Department of Hematology, Helsinki University Hospital Comprehensive Cancer Center, Helsinki, Finland. [7] Translational Immunology Research Program and Department of Clinical Chemistry, University of Helsinki, Helsinki, Finland. [8] Department of Pediatrics and Adolescent Medicine, Turku University Hospital and University of Turku, Turku, Finland. [9] Université Montpellier, Montpellier, France. [10] MACVIA-France, Montpellier, France. [11] Corporate Member of Freie Universität Berlin, Humboldt-Universität Zu Berlin, and Berlin Institute of Health, Comprehensive Allergy Center, Department of Dermatology and Allergy, Charité–Universitätsmedizin Berlin, Berlin, Germany.

Received: 12 June 2020 Accepted: 25 September 2020
Published online: 28 October 2020

References

1. Wang D-Y. Risk factors of allergic rhinitis: genetic or environmental? Ther Clin Risk Manag. 2005;1(2):115–23.
2. Fokkens WJ, Lund VJ, Mullol J, Bachert C, Alobid I, Baroody F, et al. EPOS 2012: European position paper on rhinosinusitis and nasal polyps 2012. A summary for otorhinolaryngologists Rhinology. 2012;50(1):1–12.
3. de Loos DD, Lourijsen ES, Wildeman MAM, Freling NJM, Wolvers MDJ, Reitsma S, et al. Prevalence of chronic rhinosinusitis in the general population based on sinus radiology and symptomatology. J Allergy Clin Immunol. 2019;143(3):1207–14.
4. Hirsch AG, Nordberg C, Bandeen-Roche K, Tan BK, Schleimer RP, Kern RC, et al. Radiologic sinus inflammation and symptoms of chronic rhinosinusitis in a population-based sample. Allergy. 2019;all.14106.
5. Fokkens WJ, Lund VJ, Hopkins C, Hellings PW, Kern R, Reitsma S, et al. European Position Paper on Rhinosinusitis and Nasal Polyps, vol 103. Dieudonné Nyenbue Tshipukane. 2020
6. Wikstén J, Toppila-Salmi S, Mäkelä M. Primary prevention of airway allergy. Curr Treat Options Allergy. 2018;5(4):347–55.
7. Dharmage SC, Perret JL, Custovic A. Epidemiology of asthma in children and adults. Front Pediatrics. 2019;7:246.
8. Bashiardes S, Zilberman-Schapira G, Elinav E. Use of metatranscriptomics in microbiome research. Bioinform Biol Insights. 2016;10:19–25.
9. Laulajainen-Hongisto A, Toppila-Salmi S, Luukkainen A, Kern R. Airway Epithelial Dynamics in Allergy and Related Chronic Inflammatory Airway Diseases. Front Cell Dev Biol. 2020;8(204).
10. Ober C. Asthma genetics in the post-GWAS Era. In: Annals of the American Thoracic Society. American Thoracic Society; 2016. p. S85–90.
11. Mattila P, Renkonen J, Toppila-Salmi S, Parviainen V, Joenväärä S, Alff-Tuomala S, et al. Time-series nasal epithelial transcriptomics during natural pollen exposure in healthy subjects and allergic patients. Allergy Eur J Allergy Clin Immunol. 2010;65(2):175–83.
12. Dickson RP, Erb-Downward JR, Huffnagle GB. The role of the bacterial microbiome in lung disease. Expert Review of Respiratory Medicine. 2013;7:245–57.
13. Busse WW, Lemanske RF, Gern JE. Role of viral respiratory infections in asthma and asthma exacerbations. Lancet. 2010;376:826–34.
14. Li N, Peters AT. Chronic rhinosinusitis management beyond intranasal steroids and saline solution irrigations. Allergy Asthma Proc. 2015;36(5):339–43.
15. Just J, Pierre SP, Amat F, Gouvis-Echraghi R, Lambert-Guillemot N, Guiddir T, et al. What lessons can be learned about asthma phenotypes in children from cohort studies? Pediatric Allergy Immunol. 2015;26:300–5.
16. Lemmetyinen RE, Karjalainen JV, But A, Renkonen RLO, Pekkanen JR, Toppila-Salmi SK, et al. Higher mortality of adults with asthma: a 15-year follow-up of a population-based cohort. Allergy. 2018;73(7):1479–88.
17. Hanif T, Dhaygude K, Kankainen M, Renkonen J, Mattila P, Ojala T, et al. Birch pollen allergen immunotherapy reprograms nasal epithelial transcriptome and recovers microbial diversity. J Allergy Clin Immunol. 2019;143(6):2293–2296.e11.
18. Lampi J, Koskela H, Hartikainen A-L, Ramasamy A, Couto Alves A, Järvelin M-R, et al. Farm environment during infancy and lung function at the age of 31: a prospective birth cohort study in Finland. BMJ Open. 2015;5(7):e007350.
19. Demenais F, Margaritte-Jeannin P, Barnes KC, Cookson WOC, Altmüller J, Ang W, et al. Multiancestry association study identifies new asthma risk loci that colocalize with immune-cell enhancer marks. Nat Genet. 2018;50(1):42–50.
20. Ober C, Yao T-C. The genetics of asthma and allergic disease: a 21st century perspective. Immunol Rev. 2011;242(1):10–30.
21. Oakley GM, Curtin K, Orb Q, Schaefer C, Orlandi RR, Alt JA. Familial risk of chronic rhinosinusitis with and without nasal polyposis: genetics or environment. Int Forum Allergy Rhinol. 2015;5(4):276–82.
22. Kim KW, Ober C. Lessons learned from GWAS of asthma. Allergy, asthma and immunology research. Korean Acad Asthma Allergy Clin Immunol. 2019;11:170–87.
23. Willis-Owen SAG, Cookson WOC, Moffatt MF. The genetics and genomics of asthma. Annu Rev Genomics Hum Genet. 2018;19(1):223–46.
24. Schoettler N, Rodríguez E, Weidinger S, Ober C. Advances in asthma and allergic disease genetics: Is bigger always better? J Allergy Clin Immunol. 2019;144(6):1495–506.
25. Yu G, Wang LG, Han Y, He QY. ClusterProfiler: An R package for comparing biological themes among gene clusters. Omi A J Integr Biol. 2012;16(5):284–7.
26. Chupp GL, Lee CG, Jarjour N, Shim YM, Holm CT, He S, et al. A chitinase-like protein in the lung and circulation of patients with severe asthma. N Engl J Med. 2007;357(20):2016–27.

27. Ober C, Tan Z, Sun Y, Possick JD, Pan L, Nicolae R, et al. Effect of variation in CHI3L1 on serum YKL-40 level, risk of asthma, and lung function. N Engl J Med. 2008;358(16):1682–91.
28. Lemanske RF. The Childhood Origins of Asthma (COAST) study. In: Pediatric Allergy and Immunology, Supplement. 2002. p. 38–43.
29. Hulpiau P, van Roy F. Molecular evolution of the cadherin superfamily. Int J Biochem Cell Biol. 2009;41:349–69.
30. Griggs TF, Bochkov YA, Basnet S, Pasic TR, Brockman-Schneider RA, Palmenberg AC, et al. Rhinovirus C targets ciliated airway epithelial cells. Respir Res. 2017;18(1):84.
31. Bønnelykke K, Sleiman P, Nielsen K, Kreiner-Møller E, Mercader JM, Belgrave D, et al. A genome-wide association study identifies CDHR3 as a susceptibility locus for early childhood asthma with severe exacerbations. Nat Genet. 2014;46(1):51–5.
32. Stein MM, Thompson EE, Schoettler N, Helling BA, Magnaye KM, Stanhope C, et al. A decade of research on the 17q12-21 asthma locus: Piecing together the puzzle. J Allergy Clin Immunol. 2018;142(3):749–64.e3.
33. Das S, Miller M, Broide DH. Chromosome 17q21 Genes ORMDL3 and GSDMB in Asthma and Immune Diseases. In: Advances in Immunology. Academic Press Inc.; 2017. p. 1–52.
34. Broz P, Pelegrín P, Shao F. The gasdermins, a protein family executing cell death and inflammation. Nature Reviews Immunology: Nature Publishing Group; 2019.
35. Panganiban RA, Sun M, Dahlin A, Park HR, Kan M, Himes BE, et al. A functional splice variant associated with decreased asthma risk abolishes the ability of gasdermin B to induce epithelial cell pyroptosis. J Allergy Clin Immunol. 2018;142(5):1469–78.e2.
36. Moffatt MF, Gut IG, Demenais F, Strachan DP, Bouzigon E, Heath S, et al. A large-scale, consortium-based genomewide association study of asthma. N Engl J Med. 2010;363(13):1211–21.
37. Ferreira MAR, Matheson MC, Tang CS, Granell R, Ang W, Hui J, et al. Genome-wide association analysis identifies 11 risk variants associated with the asthma with hay fever phenotype. J Allergy Clin Immunol. 2014;133(6):1564–71.
38. Torgerson DG, Ampleford EJ, Chiu GY, Gauderman WJ, Gignoux CR, Graves PE, et al. Meta-analysis of genome-wide association studies of asthma in ethnically diverse North American populations. Nat Genet. 2011;43(9):887–92.
39. Bønnelykke K, Matheson MC, Pers TH, Granell R, Strachan DP, Alves AC, et al. Meta-analysis of genome-wide association studies identifies ten loci influencing allergic sensitization. Nat Genet. 2013;45(8):902–6.
40. Ferreira MA, Vonk JM, Baurecht H, Marenholz I, Tian C, Hoffman JD, et al. Shared genetic origin of asthma, hay fever and eczema elucidates allergic disease biology. Nat Genet. 2017;49(12):1752–7.
41. Hinds DA, McMahon G, Kiefer AK, Do CB, Eriksson N, Evans DM, et al. A genome-wide association meta-analysis of self-reported allergy identifies shared and allergy-specific susceptibility loci. Nat Genet. 2013;45(8):907–11.
42. Yan Q, Brehm J, Pino-Yanes M, Forno E, Lin J, Oh SS, et al. A meta-analysis of genome-wide association studies of asthma in Puerto Ricans. Eur Respir J. 2017;49(5).
43. Poole A, Urbanek C, Eng C, Schageman J, Jacobson S, O'Connor BP, et al. Dissecting childhood asthma with nasal transcriptomics distinguishes subphenotypes of disease. J Allergy Clin Immunol. 2014;133(3):670.
44. Das S, Miller M, Beppu AK, Mueller J, McGeough MD, Vuong C, et al. GSDMB induces an asthma phenotype characterized by increased airway responsiveness and remodeling without lung inflammation. Proc Natl Acad Sci U S A. 2016;113(46):13132–7.
45. Nieuwenhuis MA, Siedlinski M, van den Berge M, Granell R, Li X, Niens M, et al. Combining genomewide association study and lung eQTL analysis provides evidence for novel genes associated with asthma. Allergy Eur J Allergy Clin Immunol. 2016;71(12):1712–20.
46. Awwad MHS, Kriegsmann K, Plaumann J, Benn M, Hillengass J, Raab MS, et al. The prognostic and predictive value of IKZF1 and IKZF3 expression in T-cells in patients with multiple myeloma. Oncoimmunology. 2018;7(10):e1486356.
47. Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, et al. Tissue-based map of the human proteome. Science. 2015;347(6220):1260419.
48. Vandevenne M, Jacques DA, Artuz C, Nguyen CD, Kwan AHY, Segal DJ, et al. New insights into DNA recognition by zinc fingers revealed by structural analysis of the oncoprotein ZNF217. J Biol Chem. 2013;288(15):10616–27.
49. Cohen PA, Donini CF, Nguyen NT, Lincet H, Vendrell JA. The dark side of ZNF217, a key regulator of tumorigenesis with powerful biomarker value. Oncotarget. 2015;6(39):41566–81.
50. Cassandri M, Smirnov A, Novelli F, Pitolli C, Agostini M, Malewicz M, et al. Zinc-finger proteins in health and disease, vol. 3. Springer Nature: Cell Death Discovery; 2017.
51. Wu AC, Himes BE, Lasky-Su J, Litonjua A, Peters SP, Lima J, et al. Inhaled corticosteroid treatment modulates ZNF432 gene variant's effect on bronchodilator response in asthmatics. J Allergy Clin Immunol. 2014;133(3):723–8.
52. O'Regan GM, Sandilands A, McLean WHI, Irvine AD. Filaggrin in atopic dermatitis. J Allergy Clin Immunol. 2009,124(3 SUPPL. 2).
53. Van Den Oord RAHM, Sheikh A. Filaggrin gene defects and risk of developing allergic sensitisation and allergic disorders: systematic review and meta-analysis. BMJ. 2009;339(7712):86–8.
54. Bin L, Leung DYM. Genetic and epigenetic studies of atopic dermatitis. Allergy, Asthma and Clinical Immunology. 2016;12:52.
55. Weidinger S, Willis-Owen SAG, Kamatani Y, Baurecht H, Morar N, Liang L, et al. A genome-wide association study of atopic dermatitis identifies loci with overlapping effects on asthma and psoriasis. Hum Mol Genet. 2013;22(23):4841–56.
56. Tsuji G, Hashimoto-Hachiya A, Kiyomatsu-Oda M, Takemura M, Ohno F, Ito T, et al. Aryl hydrocarbon receptor activation restores filaggrin expression via OVOL1 in atopic dermatitis. Cell Death Dis. 2017;8(7):e2931.
57. Hashimoto-Hachiya A, Tsuji G, Murai M, Yan X, Furue M. Upregulation of FLG, LOR, and IVL expression by rhodiola crenulata root extract via aryl hydrocarbon receptor: Differential involvement of OVOL1. Int J Mol Sci. 2018;19(6):1654.
58. De Benedetto A, Qualia CM, Baroody FM, Beck LA. Filaggrin expression in oral, nasal, and esophageal mucosa. J Investigative Dermatol. 2008;128:1594–7.
59. Renkonen J, Toppila-Salmi S, Joenväärä S, Mattila P, Parviainen V, Hagström J, et al. Expression of Toll-like receptors in nasal epithelium in allergic rhinitis. APMIS. 2015;123(8):716–25.
60. Anthoni M, Wang G, Leino MS, Lauerma AI, Alenius HT, Wolff HJ. Smad3 -signalling and Th2 cytokines in normal mouse airways and in a mouse model of asthma. Int J Biol Sci. 2007;3(7):477–85.
61. Yasukawa A, Hosoki K, Toda M, Miyake Y, Matsushima Y, Matsumoto T, et al. Eosinophils promote epithelial to mesenchymal transition of bronchial epithelial cells. Chu HW, editor. PLoS One. 2013;8(5):e64281.
62. Kitamura H, Cambier S, Somanath S, Barker T, Minagawa S, Markovics J, et al. Mouse and human lung fibroblasts regulate dendritic cell trafficking, airway inflammation, and fibrosis through integrin αvβ8 - Mediated activation of TGF-β. J Clin Invest. 2011;121(7):2863–75.
63. Fernando MMA, Stevens CR, Walsh EC, De Jager PL, Goyette P, Plenge RM, et al. Defining the role of the MHC in autoimmunity: a review and pooled analysis. PLoS Genetics. 2008;4:e1000024.
64. Fawzy MS, Elgazzaz MG, Ibrahim A, Hussein MH, Khashana MS, Toraih EA. Association of group-specific component exon 11 polymorphisms with bronchial asthma in children and adolescents. Scand J Immunol. 2019;89(3):e12740.
65. Global strategy for asthma management and prevention. (cited 2020 Apr 20). www.ginasthma.org
66. Boonpiyathad T, Sözener ZC, Satitsuksanoa P, Akdis CA. Immunologic mechanisms in asthma. Seminars in Immunology. 2019;46:101333.
67. Fahy JV. Type 2 inflammation in asthma-present in most, absent in many. Nature Rev Immunol. 2015;15:57–655.
68. Lloyd CM. IL-33 family members and asthma - bridging innate and adaptive immune responses. Curr Opin Immunol. 2010;22:800–6.
69. Takatori H, Makita S, Ito T, Matsuki A, Nakajima H. Regulatory mechanisms of IL-33-ST2-mediated allergic inflammation. Front Immunol. 2018;9:2004.
70. Cayrol C, Girard JP. IL-33: An alarmin cytokine with crucial roles in innate immunity, inflammation and allergy. Curr Opin Immunol. 2014;31:31–7.
71. Papadopoulos NG, Barnes P, Canonica GW, Gaga M, Heaney L, Menzies-Gow A, et al. The Evolving Algorithm of Biological Selection in Severe Asthma. Allergy. 2020; 14256.

72. Nadif R, Siroux V, Boudier A, Le Moual N, Just J, Gormand F, et al. Blood granulocyte patterns as predictors of asthma phenotypes in adults from the EGEA study. Eur Respir J. 2016;48(4):1040–51.
73. Préfontaine D, Nadigel J, Chouiali F, Audusseau S, Semlali A, Chakir J, et al. Increased IL-33 expression by epithelial cells in bronchial asthma. J Allergy Clin Immunol. 2010;125:752–4.
74. Smith D, Helgason H, Sulem P, Bjornsdottir US, Lim AC, Sveinbjornsson G, et al. A rare IL33 loss-of-function mutation reduces blood eosinophil counts and protects from asthma. PLoS Genet. 2017;13(3):e1006659.
75. Holgado A, Braun H, Van Nuffel E, Detry S, Schuijs MJ, Deswarte K, et al. IL-33trap is a novel IL-33–neutralizing biologic that inhibits allergic airway inflammation. J Allergy Clin Immunol. 2019;144(1):204–15.
76. Braun H, Afonina IS, Mueller C, Beyaert R. Dichotomous function of IL-33 in health and disease: From biology to clinical implications. Biochem Pharmacol. 2018;148:238–52.
77. Khurana Hershey GK. IL-13 receptors and signaling pathways: An evolving web. J Allergy Clin Immunol. 2003;111:677–90.
78. Sastre J, Dávila I. Dupilumab: A new paradigm for the treatment of allergic diseases. J Investig Allergol Clin Immunol. 2018;28:139–50.
79. Bachert C, Han JK, Desrosiers M, Hellings PW, Amin N, Lee SE, et al. Efficacy and safety of dupilumab in patients with severe chronic rhinosinusitis with nasal polyps (LIBERTY NP SINUS-24 and LIBERTY NP SINUS-52): results from two multicentre, randomised, double-blind, placebo-controlled, parallel-group phase 3 trials. Lancet. 2019;394(10209):1638–50.
80. Zhu Z, Homer RJ, Wang Z, Chen Q, Geba GP, Wang J, et al. Pulmonary expression of interleukin-13 causes inflammation, mucus hypersecretion, subepithelial fibrosis, physiologic abnormalities, and eotaxin production. J Clin Invest. 1999;103(6):779–88.
81. Zheng T, Zhu Z, Wang Z, Homer RJ, Ma B, Riese RJ, et al. Inducible targeting of IL-13 to the adult lung causes matrix metalloproteinase- and cathepsin-dependent emphysema. J Clin Invest. 2000;106(9):1081–93.
82. Accordini S, Calciano L, Bombieri C, Malerba G, Belpinati F, Lo Presti AR, et al. An Interleukin 13 polymorphism is associated with symptom severity in adult subjects with ever asthma. PLoS ONE. 2016;11(3):e0151292.
83. Ito T, Liu YJ, Arima K. Cellular and molecular mechanisms of TSLP function in human allergic disorders - TSLP programs the "Th2 code" in dendritic cells. Allergol Int. 2012;61:35–433.
84. Corren J, Parnes JR, Wang L, Mo M, Roseti SL, Griffiths JM, et al. Tezepelumab in adults with uncontrolled asthma. N Engl J Med. 2017;377(10):936–46.
85. Ray A, Cohn L. Th2 cells and GATA-3 in asthma: New insights into the regulation of airway inflammation. J Clin Investigation. 1999;104:985–93.
86. Bergqvist A, Andersson CK, Hoffmann HJ, Mori M, Shikhagaie M, Krohn IK, et al. Marked epithelial cell pathology and leukocyte paucity in persistently symptomatic severe asthma. Am J Resp Crit Care Med. 2013;188:1475–7.
87. KleinJan A, Klein Wolterink RGJ, Levani Y, de Bruijn MJW, Hoogsteden HC, van Nimwegen M, et al. Enforced expression of Gata3 in T cells and group 2 innate lymphoid cells increases susceptibility to allergic airway inflammation in mice. J Immunol. 2014;192(4):1385–94.
88. Krug N, Hohlfeld JM, Kirsten A-M, Kornmann O, Beeh KM, Kappeler D, et al. Allergen-induced asthmatic responses modified by a GATA3-specific DNAzyme. N Engl J Med. 2015;372(21):1987–95.
89. Czimmerer Z, Daniel B, Horvath A, Rückerl D, Nagy G, Kiss M, et al. The Transcription factor STAT6 mediates direct repression of inflammatory enhancers and limits activation of alternatively polarized macrophages. Immunity. 2018;48(1):75–90.e6.
90. Walford HH, Doherty TA. STAT6 and lung inflammation. JAK-STAT. 2013;2(4):e25301.
91. Förster R, Davalos-Misslitz AC, Rot A. CCR7 and its ligands: Balancing immunity and tolerance. Nat Rev Immunol. 2008;8:362–71.
92. Wei G, Jie Y, Haibo L, Chaoneng W, Dong H, Jianbing Z, et al. Dendritic cells derived exosomes migration to spleen and induction of inflammation are regulated by CCR7. Sci Rep. 2017;7:42996.
93. Li Y, Du Y, Zhang A, Jiang R, Nie X, Xiong X. Role of CCR7 on dendritic cell-mediated immune tolerance in the airways of allergy-induced asthmatic rats. Mol Med Rep. 2019;20(5):4425–32.
94. Weidinger S, Gieger C, Rodriguez E, Baurecht H, Mempel M, Klopp N, et al. Genome-wide scan on total serum IgE levels identifies FCER1A as novel susceptibility locus. PLoS Genet. 2008;4(8):e1000166.
95. Granada M, Wilk JB, Tuzova M, Strachan DP, Weidinger S, Albrecht E, et al. A genome-wide association study of plasma total IgE concentrations in the Framingham Heart Study. J Allergy Clin Immunol. 2012;129(3):840–5e21.
96. Dar SA, Rai G, Ansari MA, Akhter N, Gupta N, Sharma S, et al. FcεR1α gene polymorphism shows association with high IgE and anti-FcεR1α in Chronic Rhinosinusitis with Nasal Polyposis. J Cell Biochem. 2018;119(5):4142–9.
97. Shin JS, Greer AM. The role of FcεRI expressed in dendritic cells and monocytes. Cell Mol Life Sci. 2015;72:2349–60.
98. Ramasamy A, Curjuric I, Coin LJ, Kumar A, McArdle WL, Imboden M, et al. A genome-wide meta-analysis of genetic variants associated with allergic rhinitis and grass sensitization and their interaction with birth order. J Allergy Clin Immunol. 2011;128(5):996–1005.
99. Vogelzang A, McGuire HM, Yu D, Sprent J, Mackay CR, King C. A fundamental role for interleukin-21 in the generation of T follicular helper cells. Immunity. 2008;29(1):127–37.
100. Liao W, Lin JX, Wang L, Li P, Leonard WJ. Modulation of cytokine receptors by IL-2 broadly regulates differentiation into helper T cell lineages. Nat Immunol. 2011;12(6):551–9.
101. Liénart S, Merceron R, Vanderaa C, Lambert F, Colau D, Stockis J, et al. Structural basis of latent TGF-β1 presentation and activation by GARP on human regulatory T cells. Science. 2018;362(6417):952–6.
102. Fahey LM, Guzek R, Ruffner MA, Sullivan KE, Spergel J, Cianferoni A. EMSY is increased and activates TSLP & CCL5 expression in eosinophilic esophagitis. Pediatr Allergy Immunol. 2018;29:565–8.
103. Amaral AFS, Minelli C, Guerra S, Wjst M, Probst-Hensch N, Pin I, et al. The locus C11orf30 increases susceptibility to poly-sensitization. Allergy Eur J Allergy Clin Immunol. 2015;70(3):328–33.
104. Toppila-Salmi S, Huhtala H, Karjalainen J, Renkonen R, Mäkelä MJ, Wang DY, et al. Sensitization pattern affects the asthma risk in Finnish adult population. Allergy Eur J Allergy Clin Immunol. 2015;70(9):1112–20.
105. Asai Y, Eslami A, van Ginkel CD, Akhabir L, Wan M, Ellis G, et al. Genome-wide association study and meta-analysis in multiple populations identifies new loci for peanut allergy and establishes C11orf30/EMSY as a genetic risk factor for food allergy. J Allergy Clin Immunol. 2018;141(3):991–1001.
106. Ohanian M, Rozovski U, Kanagal-Shamanna R, Abruzzo LV, Loghavi S, Kadia T, et al. MYC protein expression is an important prognostic factor in acute myeloid leukemia. Leuk Lymphoma. 2019;60(1):37–48.
107. Ye L, Pan J, Liang M, Pasha MA, Shen X, D'Souza SS, et al. A critical role for c-Myc in group 2 innate lymphoid cell activation. Allergy Eur J Allergy Clin Immunol. 2019;75(4):841–52.
108. Volckaert T, Campbell A, De Langhe S. c-Myc regulates proliferation and Fgf10 expression in airway smooth muscle after airway epithelial injury in mouse. PLoS ONE. 2013;8(8):e71426.
109. Rogers NK, Clements D, Dongre A, Harrison TW, Shaw D, Johnson SR. Extra-cellular matrix proteins induce matrix metalloproteinase-1 (MMP-1) activity and increase airway smooth muscle contraction in asthma. PLoS ONE. 2014;9(2):e90565.
110. Zhao Y, Zhang Y, Zhang L. Variant of PBX2 gene in the 6p21.3 asthma susceptibility locus is associated with allergic rhinitis in Chinese subjects. Int Forum Allergy Rhinol. 2016;6(5):537–43.
111. Clark H, Granell R, Curtin JA, Belgrave D, Simpson A, Murray C, et al. Differential associations of allergic disease genetic variants with developmental profiles of eczema, wheeze and rhinitis. Clin Exp Allergy. 2019;49(11):1475–86.
112. Namkung JH, Lee JE, Kim E, Kim HJ, Seo EY, Jang HY, et al. Association of polymorphisms in genes encoding IL-4, IL-13 and their receptors with atopic dermatitis in a Korean population. Exp Dermatol. 2011;20(11):915–9.
113. Li X, Hawkins GA, Moore WC, Hastie AT, Ampleford EJ, Milosevic J, et al. Expression of asthma susceptibility genes in bronchial epithelial cells and bronchial alveolar lavage in the Severe Asthma Research Program (SARP) cohort. J Asthma. 2016;53(8):775–82.
114. Daya M, Rafaels N, Brunetti TM, Chavan S, Levin AM, Shetty A, et al. Association study in African-admixed populations across the Americas

recapitulates asthma risk loci in non-African populations. Nat Commun. 2019;10(1):1.

115. Acevedo N, Sääf A, Söderhäll C, Melén E, Mandelin J, Pietras CO, et al. Interaction between Retinoid Acid Receptor-Related Orphan Receptor Alpha (RORA) and Neuropeptide S Receptor 1 (NPSR1) in Asthma. PLoS ONE. 2013;8(4):e60111.
116. Pulkkinen V, Acevedo N, Söderhäll C, Melen E, Orsmark-Pietras C, Ezer S, et al. Interaction of retinoid acid receptor-related orphan receptor alpha (RORA) and neuropeptide S receptor 1 (NPSR1) in asthma. Eur Respir J. 2012;40(Suppl 56).
117. Yang IV, Pedersen BS, Liu A, O'Connor GT, Teach SJ, Kattan M, et al. DNA methylation and childhood asthma in the inner city. J Allergy Clin Immunol. 2015;136(1):69–80.
118. Sood R, Kamikubo Y, Liu P. Role of RUNX1 in hematological malignancies. Blood Am Soc Hematol. 2017;129:2070–82.
119. Toppila-Salmi S, Luukkainen A, Xu B, Lampi J, Auvinen J, Dhaygude K, et al. Maternal smoking during pregnancy affects adult onset of asthma in offspring: a follow up from birth to age 46 years. Eur Respir J. 2020;in press.
120. Gudbjartsson DF, Bjornsdottir US, Halapi E, Helgadottir A, Sulem P, Jonsdottir GM, et al. Sequence variants affecting eosinophil numbers associate with asthma and myocardial infarction. Nat Genet. 2009;41(3):342–7.
121. Hirota T, Takahashi A, Kubo M, Tsunoda T, Tomita K, Doi S, et al. Genome-wide association study identifies three new susceptibility loci for adult asthma in the Japanese population. Nat Genet. 2011;43(9):893–6.
122. Zhu Z, Lee PH, Chaffin MD, Chung W, Loh PR, Lu Q, et al. A genome-wide cross-trait analysis from UK Biobank highlights the shared genetic architecture of asthma and allergic diseases. Nat Genet. 2018;50(6):857–64.
123. Gao Y, Li J, Zhang Y, Zhang L. Replication study of susceptibility variants associated with allergic rhinitis and allergy in Han Chinese. Allergy, Asthma Clin Immunol. 2020;16(1):13.
124. Blankenbach KV, Bruno G, Wondra E, Spohner AK, Aster NJ, Vienken H, et al. The WD40 repeat protein, WDR36, orchestrates sphingosine kinase-1 recruitment and phospholipase C-β activation by G_q-coupled receptors. Biochim Biophys Acta Mol Cell Biol Lipids. 2020;158704.
125. Palmieri F. The mitochondrial transporter family SLC25: Identification, properties and physiopathology. Mol Aspects Med. 2013;34:465–84.
126. Noguchi E, Sakamoto H, Hirota T, Ochiai K, Imoto Y, Sakashita M, et al. Genome-wide association study identifies HLA-DP as a susceptibility gene for pediatric asthma in Asian populations. PLoS Genet. 2011;7(7):e1002170.
127. Ngan E, Stoletov K, Smith HW, Common J, Muller WJ, Lewis JD, et al. LPP is a Src substrate required for invadopodia formation & efficient breast cancer lung metastasis. Nat Commun. 2017;8:1.
128. Qiu W, Cho MH, Riley JH, Anderson WH, Singh D, Bakke P, et al. Genetics of Sputum Gene Expression in Chronic Obstructive Pulmonary Disease. Wurfel MM, editor. PLoS One. 2011;6(9):e24395.
129. Haque A, Koide N, Odkhuu E, Tsolmongyn B, Naiki Y, Komatsu T, et al. Mouse pyrin and HIN domain family member 1 (pyhin1) protein positively regulates LPS-induced IFN-β and NO production in macrophages. Innate Immun. 2014;20(1):40–8.
130. Wang M, Sun Y, Zheng F, Gao Y, Zhang S, Wei Y, et al. Changes of respiratory function on urine volume, renal AQP1, and AQP2 in mice with bronchial asthma and effects of lung-diffusing herb Platycodon grandiflorus (JACQ.) A.D.C. J Tradit Chinese Med Sci. 2018;5(1):72–9.
131. Hu Y, Shmygelska A, Tran D, Eriksson N, Tung JY, Hinds DA. GWAS of 89,283 individuals identifies genetic variants associated with self-reporting of being a morning person. Nat Commun. 2016;7:1.
132. Delfín DA, DeAguero JL, McKown EN. The extracellular matrix protein ABI3BP in cardiovascular health and disease. Front Cardiovasc Med. 2019;6:23.
133. Radder JE, Gregory AD, Leme AS, Cho MH, Chu Y, Kelly NJ, et al. Variable susceptibility to cigarette smoke-induced emphysema in 34 inbred strains of mice implicates Abi3bp in emphysema susceptibility. Am J Respir Cell Mol Biol. 2017;57(3):367–75.
134. Yang H, Chen W, Shi J, Huang D, Li J, Hu B, et al. Knockout of Abi3bp in mice does not affect their olfactory function, mental state and NNK-induced lung tumorigenesis. Acta Biochim Biophys Sin (Shanghai). 2016;48(9):820–6.
135. Khundadze M, Kollmann K, Koch N, Biskup C, Nietzsche S, Zimmer G, et al. A hereditary spastic paraplegia mouse model supports a role of ZFYVE26/SPASTIZIN for the Endolysosomal system. PLoS Genet. 2013;9(12):e1003988.
136. Conde L, Bracci PM, Richardson R, Montgomery SB, Skibola CF. Integrating GWAS and expression data for functional characterization of disease-associated SNPs: an application to follicular lymphoma. Am J Hum Genet. 2013;92(1):126–30.
137. Bunyavanich S, Schadt EE, Himes BE, Lasky-Su J, Qiu W, Lazarus R, et al. Integrated genome-wide association, coexpression network, and expression single nucleotide polymorphism analysis identifies novel pathway in allergic rhinitis. BMC Med Genomics. 2014;7(1):48.
138. Kristjansson RP, Benonisdottir S, Davidsson OB, Oddsson A, Tragante V, Sigurdsson JK, et al. A loss-of-function variant in ALOX15 protects against nasal polyps and chronic rhinosinusitis. Nat Genet. 2019;51(2):267–76.
139. Lemonnier N, Melén E, Jiang Y, Joly S, Ménard C, Aguilar D, et al. A novel whole blood gene expression signature for asthma, dermatitis, and rhinitis multimorbidity in children and adolescents. Allergy Eur J Allergy Clin Immunol. 2020. https://doi.org/10.1111/all.14314.
140. Comer DM, Elborn JS, Ennis M. Comparison of nasal and bronchial epithelial cells obtained from patients with COPD. PLoS ONE. 2012;7(3):e32924.
141. Pandey G, Pandey OP, Rogers AJ, Ahsen ME, Hoffman GE, Raby BA, et al. A nasal brush-based classifier of asthma identified by machine learning analysis of nasal RNA sequence data. Sci Rep. 2018;8(1):1.
142. Edris A, De Feyter S, Maes T, Joos G, Lahousse L. Monoclonal antibodies in type 2 asthma: a systematic review and network meta-analysis. Respir Res. 2019;20:179.
143. Olafsdottir TA, Theodors F, Bjarnadottir K, Bjornsdottir US, Agustsdottir AB, Stefansson OA, et al. Eighty-eight variants highlight the role of T cell regulation and airway remodeling in asthma pathogenesis. Nat Commun. 2020;11(1):1.
144. Duvall MG, Krishnamoorthy N, Levy BD. Non-type 2 inflammation in severe asthma is propelled by neutrophil cytoplasts and maintained by defective resolution. Allergol Int. 2019;68:143–9.
145. Revez JA, Bain L, Chapman B, Powell JE, Jansen R, Duffy DL, et al. A new regulatory variant in the interleukin-6 receptor gene associates with asthma risk. Genes Immun. 2013;14(7):441–6.
146. Anwar MA, Shah M, Kim J, Choi S. Recent clinical trends in Toll-like receptor targeting therapeutics. Med Res Rev. 2019;39:1053–90.
147. Zhu Z, Guo Y, Shi H, Liu CL, Panganiban RA, Chung W, et al. Shared genetic and experimental links between obesity-related traits and asthma subtypes in UK Biobank. J Allergy Clin Immunol. 2020;145(2):537–49.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Cell. Mol. Life Sci. (2012) 69:1917–1929
DOI 10.1007/s00018-011-0886-0

Cellular and Molecular Life Sciences

REVIEW

Deciphering the structure and function of FcεRI/mast cell axis in the regulation of allergy and anaphylaxis: a functional genomics paradigm

**Jayapal Manikandan · Narasimhan Kothandaraman ·
Manoor Prakash Hande · Peter Natesan Pushparaj**

Received: 27 April 2011 / Revised: 27 October 2011 / Accepted: 7 November 2011 / Published online: 7 December 2011
© Springer Basel AG 2011

Abstract Allergy and anaphylaxis are inflammatory disorders caused by immune reactions mainly induced by immunoglobulin-E that signal through the high-affinity FcεRI receptor to release the inflammatory mediators from innate immune cells. The FcεRI/mast cell axis is potently involved in triggering various intracellular signaling molecules to induce calcium release from the internal stores, induction of transcription factors such as NF-kB, secretion of various cytokines as well as lipid mediators, and degranulation, resulting in the induction of allergy and anaphylaxis. In this review, we discuss various cellular and molecular mechanisms triggered through FcεRI/mast cell axis in allergy and anaphylaxis with a special emphasis on the functional genomics paradigm.

Keywords Allergy · Anaphylaxis · FcεRI · IgE · Cytokines · Transcription · Factors · Functional · Genomics

Introduction

The term *allergy* was first used in 1906 by Dr. Von Pirquet, an Austrian physician, to define an Immunoglobulin E (IgE)-mediated or type I hypersensitivity reaction triggered by allergens, which are antigens that bring out an allergic reaction [1]. In 1902, Richet and Portier [2] found that a dog tolerated the administration of sea anemone toxin initially; however, it died within few minutes of administering the second dose many weeks later. They coined the term *anaphylaxis*, which in Greek, *ana* means against; *phylax* means guard or protection. Richet was later awarded the Nobel Prize in Medicine and Physiology in 1913 for his work in anaphylaxis. At present, anaphylaxis refers to a life-threatening IgE-mediated allergic condition characterized by multiple organ involvement and quick onset. Drugs, insect stings, latex, and foods are the common allergens that trigger anaphylactic reactions. Very rarely, death due to anaphylaxis typically results from acute respiratory or heart failure.

Mast cells play a major effector role in the initiation as well as the propagation of various inflammatory disorders such as asthma, psoriasis, arthritis, etc. [3] and are the central effector cells in the mediation of allergy and anaphylaxis [4]. IgE-dependent or IgE-independent mast cell mediated response leads to the activation, synthesis and release of inflammatory mediators in allergy, and anaphylaxis [5–8]. In the past decade, scientists have been applying the microarray technology to explore the fundamental genetic causes of inflammatory conditions [9–13].

J. Manikandan · M. P. Hande (✉)
Department of Physiology, Yong Loo Lin School of Medicine,
National University of Singapore, Singapore 117597, Singapore
e-mail: phsmph@nus.edu.sg

J. Manikandan · N. Kothandaraman · P. N. Pushparaj
Center of Excellence in Genomic Medicine Research (CEGMR),
King Fahad Medical Research Center, King AbdulAziz
University, Jeddah 21589, Saudi Arabia

M. P. Hande
Tembusu College, University Town, 28 College Avenue East,
B1-01, Singapore 138598, Singapore

P. N. Pushparaj (✉)
Institute of Immunity, Infection and Inflammation,
College of Medicine, Veterinary and Life Sciences,
120 Glasgow Biomedical Research Centre,
University of Glasgow, Glasgow G12 8TA, UK
e-mail: peter.n.pushparaj@gmail.com;
peter.pushparaj@glasgow.ac.uk

Such transcriptomic profiling immensely helps in our critical understanding of the global effects driven by the physiological “passive sensitization” or “active stimulation” of human mast cells to design novel therapeutic strategies to ameliorate and effectively manage both allergy and anaphylaxis.

Immunoglobulins and IgE

Defense against foreign bodies and microorganisms requires the interaction between innate immunity and acquired immunity. Generally, innate and acquired responses are regulated by humoral and cell-mediated responses. The humoral responses are mediated by immunoglobulins or antibodies, which are secreted by B lymphocytes. The immunoglobulins contain a variable Fab portion in the light chains and a constant Fc portion in the heavy chains. The Fab region interacts with a specific antigen and the Fc region binds to effector molecules such as Fc receptors and complement proteins. The ability of immunoglobulins to distinguish between a variety of antigens is due to the result of somatic recombination events, which result in the alteration of their variable regions. Based on the constant Fc region, immunoglobulins are subdivided into five major classes, which includes: IgG, IgM, IgA, IgE, and IgD. The effector molecules, which distinguish these immunoglobulins, are primarily Fc receptors, transport receptors and complement [14]. Every major immunoglobulin class has its own specific Fc receptor (FcR): FcαR binds IgA, FcεR binds IgE, FcγR binds IgG, FcδR binds IgD, and FcμR binds IgM [15–18].

IgE receptors and nomenclature

FcεRI-mediated reactions are the most prevalent form of immune-related diseases among humans. FcεRs play core function in the development of allergic inflammation. There are two types of IgE receptors: the high-affinity receptor (FcεRI) and the low-affinity IgE receptor (FcεRII; CD23) [19]. FcεRI is expressed on mast cells and basophils and on antigen-presenting cells whereas, FcεRII, a Ca-dependent lectin, is expressed on the surface of B cells, as well as other hematopoietic cells including Langerhans cells, macrophages, monocytes, eosinophils, and platelets. CD23, existing as CD23a and CD23b, has been associated with facilitating antigen presentation; it also negatively regulates IgE synthesis; and transports IgE-antigen complexes across epithelium [20]. Both IgE and IL-4 are shown to upregulate expression of CD23 [21].

Unbound FcεRI on the mast cell surface has a half-life of 24 h in vitro. On the other hand, FcεRI bound to IgE seem to be expressed throughout the life of the cell [22]. Aggregation of FcεRI by multivalent antigen results in several downstream intracellular signaling events linked with mast cell or basophil activation [23, 24]. When an allergen interacts with IgE that attached to mast cells or basophils by the α chain of the high-affinity IgE receptor, it will result in the activation of mast cells and release of both stored, as well as newly synthesized mediators (FcεRI*a*) [25]. Release of these mediators results in local vasodilatation, smooth muscle contraction, increased vascular permeability, and the initiation of inflammation. Such hypersensitivity reactions result in clinical manifestations including allergic rhinitis, atopic dermatitis, and anaphylaxis. In antigen-presenting cells (APCs), these receptors facilitate the IgE-mediated trapping and introduction of allergen to T cells [26, 27]. Eosinophils also have FcεRIa, but apart from activation and degranulation, it may aid in regulating local levels of IgE [28].

FcεRI: structure

FcεRI has a K_d of over 10^{-10} M binding affinity with monomeric IgE, which is the strongest when compared to the rest of the other Fc receptors for their ligands [29]. The classical form of FcεRI is tetrameric ($\alpha\beta\gamma_2$): constitutively expressed on mast cells and basophils in humans, while the trimeric form ($\alpha\gamma_2$) is present on the APCs such as monocytes, dendritic cells and Langerhans cells. The classical form consists of an IgE-binding α chain, a β chain, and a homodimeric γ chain. The β chain in FcεRI enhances receptor maturation [30] and the γ-homodimer enhances signal transduction [31]. Studies show that even in the absence of a β chain, the trimeric form of FcεRI possesses complete function [25, 32]. Both the β and γ chains are required for the efficient cellular expression of the FcεRI-α-chain [33]. In addition, studies have shown that $\alpha\gamma_2$ complexes were translocated to the periphery of the cell, meaning that the human γ chain by itself is sufficient to fight against endoplasmic reticulum retention signals in the α chain [33].

FcεRI effector cells: mast cells and basophils

Mast cells have been considered as the most vital effector cell type for allergic conditions including, to a lesser extent, basophils and neutrophils [34, 35]. The progenitors for mast cells migrate into the peripheral tissue and undergo differentiation to become mature mast cells in situ. They are mainly located within blood vessels and the epithelial lining. Mast cells of hematopoietic origin respond to signals of both innate and acquired immune response with immediate and delayed release of inflammatory mediators. Based on their site of location and granule contents, mast cells are divided into connective tissue mast cells and mucosal mast cells. Human connective tissue mast cells are observed in the skin and intestinal sub-mucosa, and their granules comprises tryptase and chymase. Studies have also shown that the granules of tryptase were found in the mast cells located in intestinal mucosa and alveoli [36]. Mast cells are central to the pathogenesis of type I hypersensitivity and mastocytosis. Mast cells are also entailed with self-responses to pathogens, autoimmune diseases, and fibrosis. Eicosanoids and cytokines are synthesized by mast cells and contribute to inflammation. Mast cells can be specifically stained with basic dyes such as toluidine blue [37].

Basophils also express high levels of FcεRI. They are derived from CD34$^+$ hematopoietic progenitor cells and play a vital role in the host defense against parasitic infections, as well as mediating type I hypersensitivity reactions. Basophils are present in blood as mature forms. However, they can be recruited to the site of inflammation where antigens are located. Cytoplasmic granules in the basophils comprise pro-inflammatory mediators. Because basophils share structural and functional similarities with mast cells, they trigger comparable effector responses as observed in mast cells upon aggregation of FcεRI. Apart from basophils and mast cells, FcεRI is also found to be expressed in low levels on eosinophils, monocytes, platelets, dendritic cells, and Langerhans cells [25].

Mast cells: morphology

Mast cells are up to 20 μm in diameter, are ovoid or irregularly elongated in shape [38]. Mast cells contain ample metachromatic cytoplasmic granules and they can be stained because of large sulfated proteoglycans in the granules. Tryptase staining detects all mast cell types and is the principal method for detecting tissue mast cells. The granule contents are crystalline under the electron microscope, but turn formless upon activation and before degranulation. Mast cells express several receptors including IL-3R, IL-4R, IL-5R, IL-9R, IL-10R, GM-CSFR, IFN-γR, C3a and C5a receptors, CCR3, CCR5, CXCR2, CXCR4, nerve growth factor receptor, Toll-like receptors (TLRs), and ST2 [39, 40].

FcεRI-mediated signal transduction

Interpreting the intracellular signaling cascade mediated by mast cell activation has major therapeutic implications for inflammatory conditions. FcεRI aggregation and mast cell activation has long been known to be an important incident in allergic conditions [41–43]. Numerous studies have focused on the intracellular signaling cascade mediated by mast cell activation and release of pro-inflammatory mediators [25, 44, 45]. Figure 1 shows the schematic representation of FcεRI-mediated intracellular signaling events. Following cross-linking of FcεRI, intracellular signal transduction is initiated by a tyrosine kinase, named Lyn, which is constitutively linked with the FcεRI receptor [46, 47]. Subsequently, Lyn phosphorylates the immunoreceptor tyrosine-based activation motif (ITAM) of the subunits of the FcεRI, leading to the activation and binding of protein tyrosine kinases (PTK) with Src homology domain 2 (SH2). One such PTK is spleen tyrosine kinase (Syk) [48]. An activated Syk subsequently leads to further tyrosine phosphorylations of several "adapter" proteins including that of linker for activation of T cells (LAT) as well as phosphorylation of phospholipase C γ (PLC-γ).

LAT phosphorylation leads to the activation of JNK and ERK; and the synthesis and release of cytokines and arachidonic acid (AA) metabolism result in target organs to cause the clinical syndrome of anaphylaxis [49]. On the other hand, phosphorylated PLC-γ yields diacylglycerol (DAG) and inositol-1,4,5-triphosphate (IP$_3$) from membrane phospholipids. DAG successively activates protein kinase C (PKC), which results in the exocytosis of stored granules and cytokine release.

Furthermore, Bruton's tyrosine kinase (Btk) is membrane targeted and activated by the binding of phosphatidylInositol-3, 4, 5-trisphosphate (PIP3) moiety to its Pleckstrin-Homology (PH) domain. Btk then phosphorylates and activates PLC-γ [50]. On the other hand, the induction of PI3 K by Syk results in the production of micromolar levels of PIP2 and PIP3. Besides, PI3 K is essential for the induction of intracellular Ca^{2+} release as well as mobilization across plasma membranes. Both PI3 K and PLCγ act synergistically on the common substrate PIP2 to produce PIP3 and IP$_3$, respectively. The IP$_3$ activates Ca^{2+} channels at different cellular compartments, primarily in the endoplasmic reticulum [ER], to release Ca^{2+} required for optimal physiological responses. As a result, the depletion of Ca^{2+} in the internal stores causes the influx of Ca^{2+} from the extracellular space through the activation of a type of store operated Ca^{2+} channel (SOCC) termed as Ca^{2+} release activated Ca^{2+} channel (I_{CRAC}) [50, 51, 52].

Studies have shown that cross-linking of FcεRI in rat basophilic leukemia cells activates sphingosine kinase, which produces sphingosine-1-phosphate (S1P). S1P is a potent sphingolipid mediator, and acts as a second messenger in releasing intracellular calcium from internal stores. This process is carried through the inositol 1,4,5-triphosphate (IP$_3$)-independent pathway [53]. Studies report that aggregation of FcεRI in human bone marrow-derived mast cells resulted in the activation of PLD1, leading to downstream activation of SPHK1 [54]. This process results in the initial release of intracellular calcium as well as degranulation of mast cell(s). Another interesting

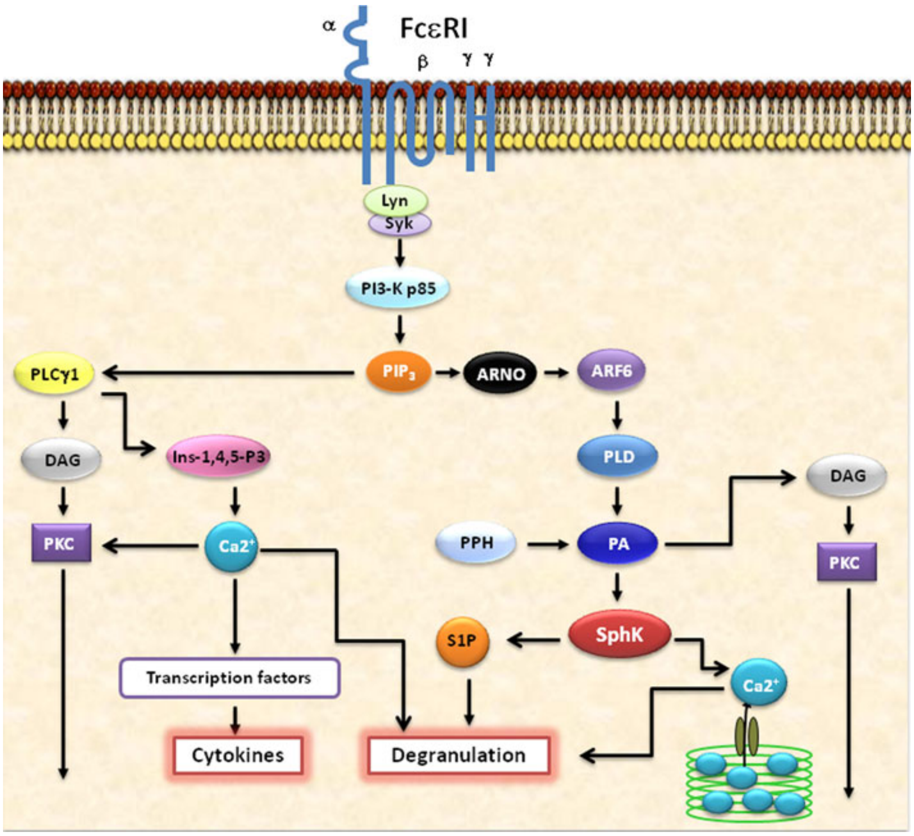


Fig. 1 Schematic representation of FcεRI-mediated intracellular signaling events in mast cells and basophils. The allergen cross linking of FcεRI receptors initiate key cellular signaling events such as the activation of PLCγ, PKCs, SPHKs, PI3Ks, etc. An important role for these major pathways in immune cell function is predicted based on their ability to regulate the transcriptional activity of cytokine genes, calcium release from the internal stores which is essential for the degranulation and further intracellular mediation of signaling processes, and the production of lipid mediators such as prostaglandins and leukotrienes. The specific blocking of these key intracellular signaling molecules such as PLCγ, PKCs, SPHKs, S1P, PI3Ks, etc., by specific chemical inhibitors or siRNAs, or therapeutic Intrabodies would certainly be of immense use in the effective management of allergy and anaphylaxis in the near future

study reports that blocking SPHK activity forbids FcεRI-mediated internalization of S1P receptors and significantly reduces degranulation [55]. Also, tyrosine kinase Lyn has been shown to be associated with recruitment and activation of SPHK to FcεRI [56].

FcεRI: regulation

In 1978, Malveaux et al. [57] reported for the first time that the presence of circulating monomeric IgE could raise FcεRI cell surface expression. Many studies have been published in the 1990s showing IgE drives FcεRI expression in both human and murine mast cells, as well as basophils [58–60]. These reports have helped in understanding the mechanisms of IgE sensitization, during which FcεRI is over-expressed at cell surface, and FcεRIs bound to IgE are activated when re-exposed to allergens. FcεRI expression as well as mast cell activation upon cross-linking IgE-bound FcεRI with polyvalent antigen lead to degranulation [34, 61]. Such IgE-mediated mast cell recognition of multivalent antigen and the subsequent intracellular signaling events are well studied in various inflammatory models [62–65].

IgE-mediated FcεRI expression was also observed in monocytes [66, 67]. However, monocytes do not express FcεRIβ and are shown to express low levels of FcεRI. FcεRIβ appears to be playing a key role in the receptor expression and has also been shown to be associated with atopic diseases [68]. A single-nucleotide polymorphism (SNP) in the FcεRIβ gene has been connected to higher IgE serum levels and result in the increased FcεRI surface expression [69]. In IgE-mediated FcεRI regulation, over-expression of receptor is caused by the interaction of IgE with the receptor [70]. The total FcεRIα content increases with surface receptor up-regulation. On the other hand, the removal of IgE results in the reduction in the receptor expression, as well as FcεRIα content [71]. Though the actual mechanism still remains unclear, IgE-FcεRI binding enhances the surface expression of FcεRI [72]. Reports show that FcεRI receptor expression in mast cells are mediated by IL-4 [60, 73]. On the whole, the factors that mainly affect the expression of FcεRI are the expression of FcεRIβ and the serum levels of IgE [60, 73].

FcεRI: biological functions

FcεRI triggers IgE-mediated activation and degranulation of mast cells and basophils. Apart from mediating type I hypersensitivity reactions, FcεRI plays other biological roles, too. First of all, FcεRI and IgE are vital for the self-defense against parasitic infections [74]. In vivo studies showed that mice lacking α chain of FcεRI were also protected against parasitic infections like the normal mice. However, it is not very clear why and how the absence of functional FcεRI/mast cell axis protects these mice against parasitic infections [74]. Interestingly, very recent study has partially answered the above statement by showing that the IgG/neutrophil axis is also key for the triggering of passive and active systemic anaphylactic shock in mice [35].

The other biological role of FcεRI is their expression on APCs such as dendritic cells and monocytes [75]. FcεRI is expressed as $\alpha\gamma_2$ form on the APCs. It is known that the $\alpha\gamma_2$ structure may assist in targeting antigen-IgE-FcεRI complexes to the intracellular antigen-presenting compartment. The IgE-dependent type of antigen presentation will make sure that it will only amplify the immune responses, which were already mediated by IgE. This type of antigen presentation can mediate cytokine production from APCs and thus, helps to modulate allergic inflammation [20, 25].

Activation of mast cells and release of mediators

The mediators released by the mast cells can be categorized into three main groups: preformed granule content, membrane derived lipid mediators, and cytokines [76–78]. Figure 2 shows the schematic representation of major types of mediators released by the mast cells. Histamine is the most clinically evident preformed granule mediators responsible for the acute symptoms. Histamine is shown to be associated with immediate type I hypersensitivity reactions [79]. The level of histamine that is produced and stored in mast cells is roughly 1 or 2 pg/cell. Histamine receptors are further classified into H_1, H_2, and H_3. H_1 receptor stimulation results in bronchial smooth muscle contraction, increased vascular permeability, nasal mucus secretion, and increased neutrophil and eosinophil chemokinesis and chemotaxis. On the other hand, H_2 receptor stimulation results in ventricular and atrial contraction, gastric acid production, airway mucus secretion, and vascular permeability as well as inhibition of basophil histamine release. Finally, H_3 receptors, which are found to be in neurons and peripheral tissues, regulate the secretion and synthesis of histamine.

Membrane-derived lipid mediators and cytokines are synthesized upon activation of mast cells. Prostaglandin D_2 (PGD_2) is the key arachidonic acid (AA) lipid metabolite released upon mast cell activation. PGD_2 is synthesized from AA by the cyclooxygenase pathway and is responsible for causing bronchospasm, hypotension, and inhibition of platelet aggregation. PGD_2 is roughly over 30 times more potent than histamine, especially in causing bronchoconstriction. Leukotrienes are slow-responding substances of anaphylaxis. Leukotrienes—LTB_4, LTC_4, LTD_4, and LTE_4 are synthesized from AA through the lipoxygenase pathway. Leukotrienes have been shown to be linked in increased vascular permeability, increased mucous gland production and cholinergic-independent bronchospasm. Leukotrienes have a slow onset but are 10–1,000 times added potent, compared to histamine in developing bronchoconstriction when dispensed by aerosol [80]. The other most potent compound and unstored phospholipid is platelet-activating factor (PAF). PAFs are known to cause aggregation of human platelets and release of platelet-derived vasoactive mediators. PAF has been implicated with anaphylaxis, including pulmonary edema and coronary vasoconstriction [81, 82]. Studies show that the blockade of PAF with inhibitors result in improved cardiac function, suggesting a key role for PAF in cardiac dysfunction [83]. Furthermore, TNF-α is a major cytokine that is in both stored as well as synthesized forms [84]. It over-expresses cell adhesion molecules, increases bronchial responsiveness, and also possesses antitumor effects. Other cytokines produced by mast cells include IL-1, IL-3, IL-4, IL-5, IL-6, and IL-8 [39].

Mast cells in acute and chronic allergic reactions

Mast cells play a central role in the initiation and development of atopic allergic reactions. The intracellular signaling events have been broadly studied on mast cells not only to know their physiological roles but also to find potential therapeutic targets. Type I hypersensitivity is the fundamental of acute allergic conditions [85]. It is stimulated by molecules released by mast cells when an allergen or antigen cross-react with membrane-bound IgE.

Allergic conditions broadly consist of two phases: sensitization and elicitation phase [86]. In the self, sensitization occurs upon the initial exposure to an allergen or antigen. Dendritic cells and macrophages are the first line of defense against antigens through phagocytosis and cell-mediated uptake of antigens. The antigen is degraded by the antigen-processing pathway and subsequently

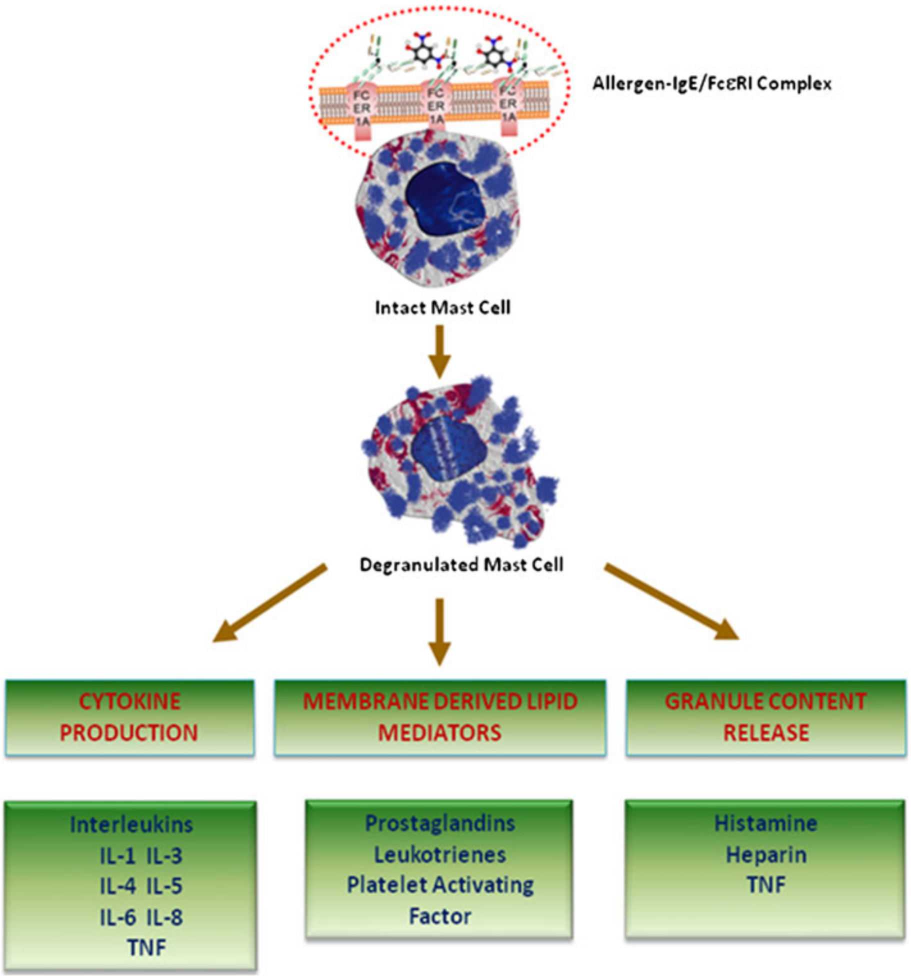


Fig. 2 The key mediators of FcεRI/Mast cell axis in allergy and anaphylaxis. The schematic representation illustrates some of the key mediators released by the mast cells upon stimulation during allergy and anaphylaxis. The mediators released by the mast cells can be categorized into three main groups: preformed granule content, membrane-derived lipid mediators, and cytokines

presented as a component of the MHC class II complexes [87]. Then the mature dendritic cells stimulate naive T cells to differentiate into either T_H1- or T_H2-type $CD4^+$ T cells [88], which aids in the maturation of B-cells to IgE secreting plasma cells. FcεRI aggregation triggers the release of pre-formed mediators, production of lipid mediators and cytokines. In an atopic condition, exposure of the nose, skin, or airway to an initial dose of allergen triggers a cutaneous reaction, results runny nose, sneezing, and wheezing in minutes. In case of asthma, these mediators develop acute reactions, including mucus production and smooth muscle contraction. This depends upon the type I immediate allergic reactions and subsequently followed by a chronic allergic reaction, which peaks 6 to 9 h post-initial exposure [89]. The cellular pathways leading to acute and chronic allergic reactions are shown in Fig. 3.

A T_H2 response has been observed in an atopic condition and IgE production will result in binding with FcεRI on the effector cell. This will in turn increase the efficacy during re-exposure to the same allergen [90–92]. During the late-phase reaction of the skin, neutrophils and eosinophils accumulate, followed by $CD4^+$ T cells and basophils that infiltrate to the site of inflammation [93]. The late-phase asthmatic condition also shares a similar pattern of cellular infiltration [94–96]. However, basophils are not mainly infiltrated into the lower airways [97, 98]. Studies also show that further amplifications of chronic allergic reactions may be mediated by histamine-releasing factors [99, 100]. Cross-linking of mast-cell-bound IgE with an antibody against IgE elicits both type I and late-phase reactions [101–104]. The immediate and delayed phases of bronchial hypersensitivity are initiated by mast cells are shown in both humans, as well as in mouse models of asthma [105, 106]. It has been reported that mast cells may also be associated with regulating the early stages of autoimmunity, especially in [107–109] diseases where auto-antibodies play a vital role [110].

High-throughput functional genomics paradigm

The latest developments of functional genomics tools, such as genome-wide transcriptomic analysis have revolutionized the approaches to answer complex scientific problems.

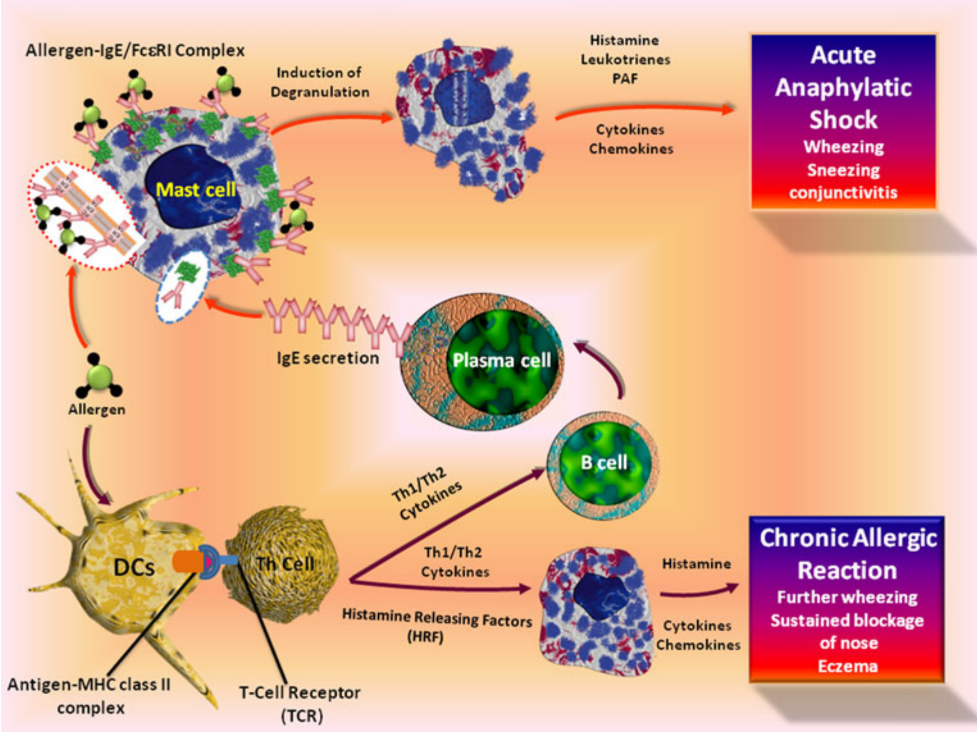

Allergen-IgE/FcεRI Complex
Induction of Degranulation
Mast cell
Histamine
Leukotrienes
PAF
Cytokines
Chemokines
Acute Anaphylatic Shock
Wheezing
Sneezing
conjunctivitis
IgE secretion
Plasma cell
Allergen
B cell
Th1/Th2 Cytokines
Th1/Th2 Cytokines
DCs
Th Cell
Histamine Releasing Factors (HRF)
Histamine
Cytokines
Chemokines
Chronic Allergic Reaction
Further wheezing
Sustained blockage of nose
Eczema
Antigen-MHC class II complex
T-Cell Receptor (TCR)

Fig. 3 The cellular pathways leading to acute and chronic allergic reactions. The binding of allergen with IgE/FcεRI complex on mast cells triggers the release of pre-formed mediators, production of lipid mediators, and cytokines leading to acute allergic conditions. Allergic conditions broadly consist of two phases: sensitization and elicitation phase. In the self, sensitization occurs upon the initial exposure to an allergen or antigen. Dendritic cells and macrophages are the first line of defense against antigens through phagocytosis and cell-mediated uptake of antigens. The antigen is subsequently degraded as a component of the MHC class II complexes. Then the mature dendritic cells stimulate naive T cells to differentiate into either T_H1- or T_H2-type $CD4^+$ T cells, which aids in the maturation of B-cells to IgE-secreting plasma cells. Chronic allergic reactions, including the late-phase reaction, may rely on a combination of events and the release of mast-cell products by histamine-releasing factors from T-helper cells

Table 1 Overview of gene expression studies performed using mast cell lines and primary mast cells

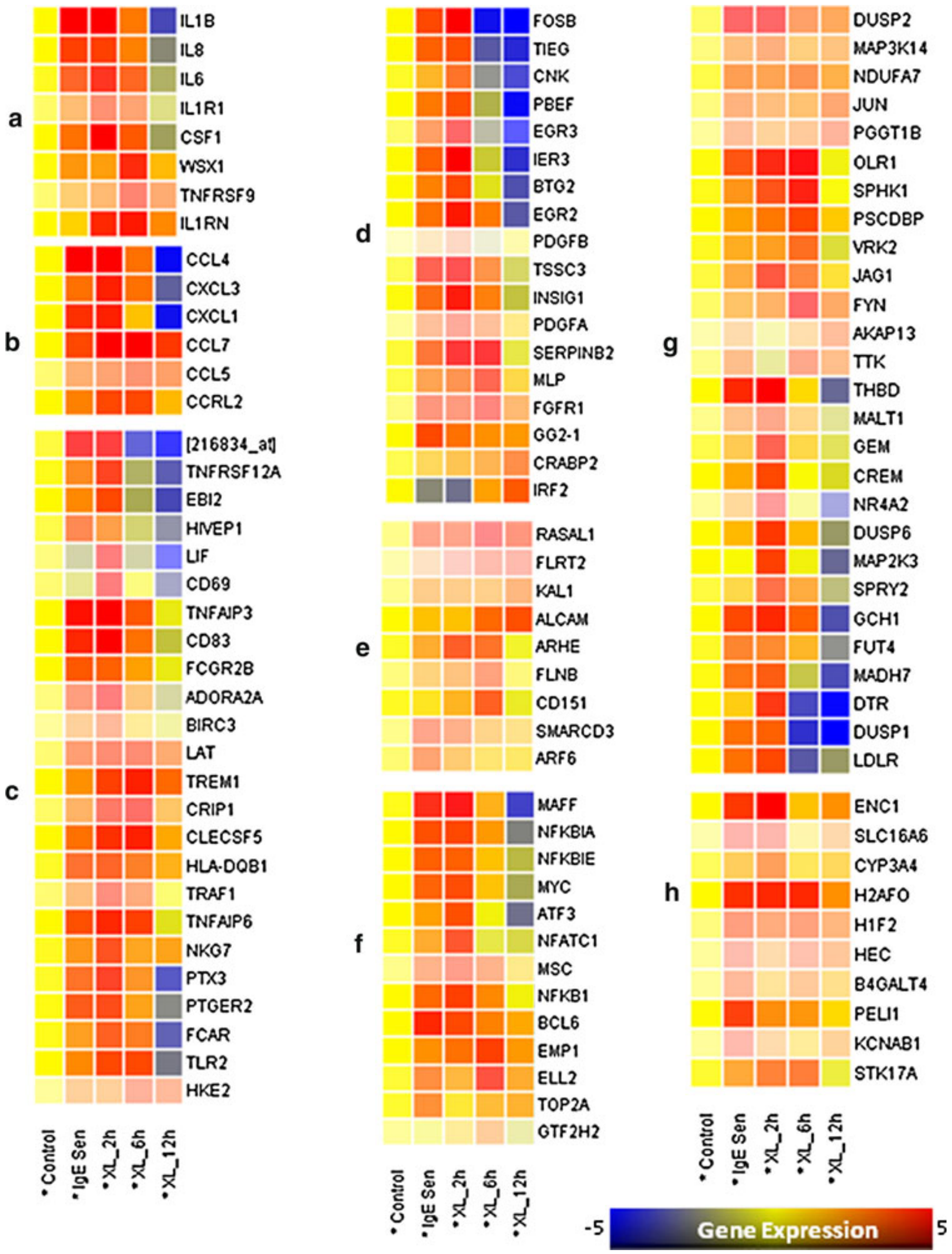
Study	Source	Stimulation	Platform	No. of genes
[116]	HUCBMCs	No stimuli	RT-PCR	32 genes
[117]	HUCBMCs	No stimuli	SAGE, RT-PCR	9,000 tags- SAGE
[118]	Human PBMNC-MCs	IgE + Anti IgE	ELISA, RT-PCR	Cytokines and chemokines
[119]	HUCBMCs, eosinophils, neutrophils and PBMNC	No stimuli	ONT array	6,000 probes
[120]	RBL-2H3	IgE + DNP	ONT array	8,799 probes
[121]	BM-mouse MCs	IgE + DNP	ELISA	Cytokines
[122]	Human lung MCs	TLR4	QPCR, ELSIA	Cytokines
[123]	Human Skin MCs	IgE or PMA	ELISA, RT-PCR	Cytokines
[124]	HUCBMCs	IgE	ELISA	IL-8 & MCP-1
[115]	HUCBMCs	IgE	ONT array	8,763 probes

The above evidence table depicts the authors of the study, source of mast cells used, stimulation method for activation of mast cells, platforms used for gene expression profiling, and number of genes screened.

MC mast cell, *HUCBMCs* human umbilical cord blood-derived mast cells, *PBMNC* peripheral blood mononuclear cell culture, *RBL-2H3* rat basophilic leukemia cells

Several groups have carried out low, medium, and high-throughput transcriptomic analysis of mast cell-mediated expression profiling, furnishing some novel pathogenic mechanisms of mast cell-mediated allergic responses. Table 1 shows various gene expression studies carried out on mast cells. The impact of these studies confirms that the mast cell stimulation and activation seem to be associated with number of pathologies including allergy and other

Fig. 4 Gene expression pattern in human mast cells following IgE sensitization and FcεRI aggregation. The raw Affymetrix GeneChip data was downloaded from gene expression omnibus (GSE1933) and normalized with parametric test based on cross gene error model (PCGEM) and subjected to one-way ANOVA and Bonferroni-Hochberg FDR ($p < 0.05$) using Genespring 7.3 The differentially expressed genes by the IgE sensitization and FcεRI cross-linking for different time points (2, 6, and 12 h) were then classified and clustered based on gene ontology (GO). Analysis to decode the differentially expressed genes implicated in biological processes such as **a** cytokines and cytokine receptors, **b** chemokines and chemokine receptors, **c** other immunoregulatory genes, **d** cell proliferation and apoptosis, **e** adhesion and cytoskeleton remodeling, **f** transcription factors and regulators, **g** signal transduction, and **h** other genes [115]



immune/inflammatory conditions. A number of studies on immune/inflammatory response show the usefulness of transcriptomic analysis to understand the underlying molecular mechanisms [11, 111–114]. With transcriptomic analysis coupled with stringent statistical measures, we show the genome-wide expression patterns of differentiated human mast cells stimulated by IgE sensitization, and FcεRI aggregation at different time points. In our earlier study, we have elucidated the molecular events which the differentiated human mast cells go through upon IgE sensitization and after complete activation in a genome-wide manner [115]. Generation of a wide-variety of cytokines and chemokines upon IgE-sensitization and FcεRI aggregation on mast cells suggests that they could be the key regulators of the immune response (Fig. 4). It may also lead to the recruitment and activation of other effector cells to the site of inflammation, which may further enhance the progression of immune response [115].

FcεRI/mast cell axis: pathway analysis of differentially expressed genes

Representation of specific inflammatory and immunoregulatory pathways among the differentially expressed genes [115] was analyzed using Pathway Studio® software

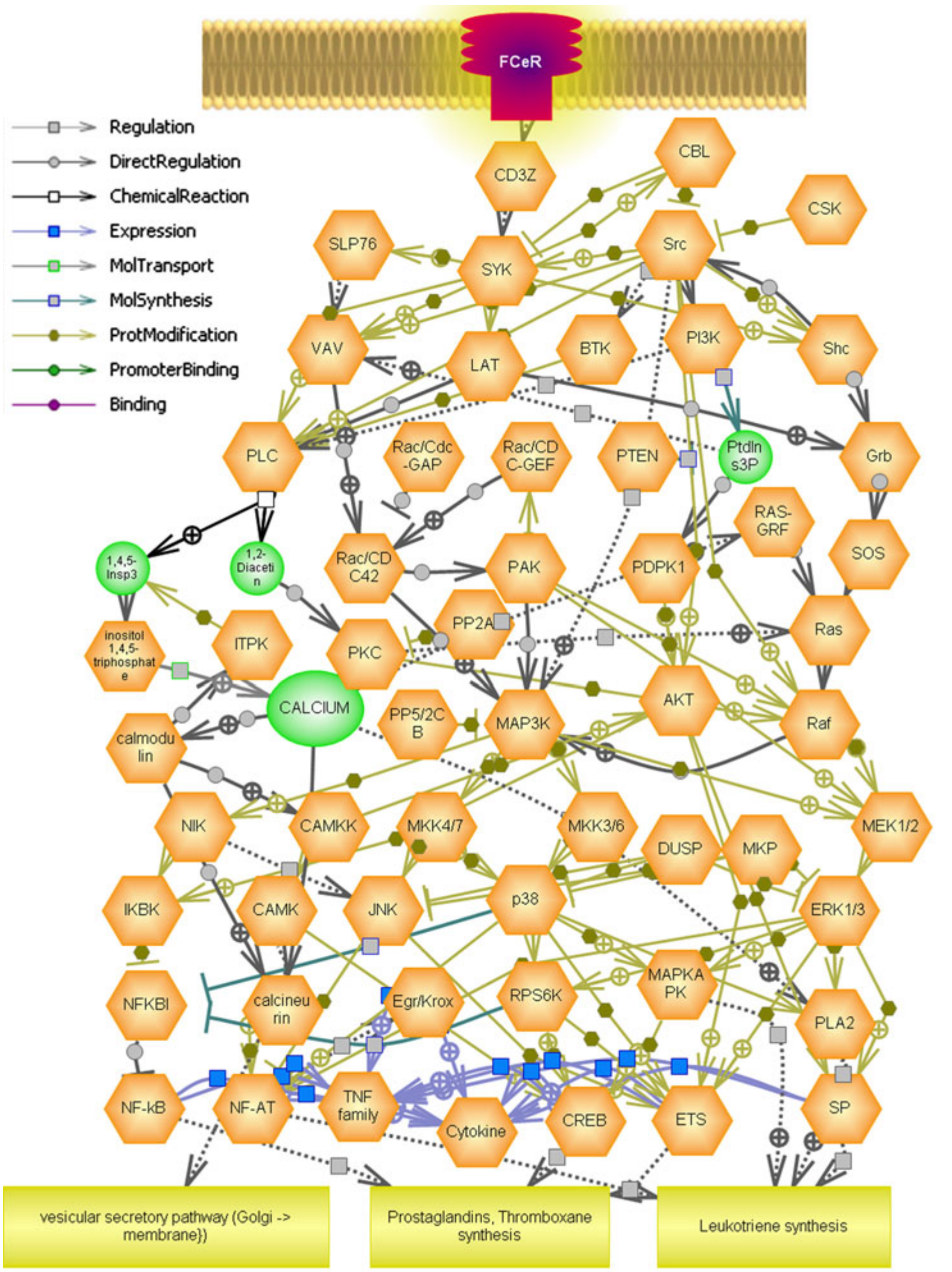

FCeR
Regulation
DirectRegulation
ChemicalReaction
Expression
MolTransport
MolSynthesis
ProtModification
PromoterBinding
Binding
CD3Z
CBL
CSK
SLP76
SYK
Src
VAV
LAT
BTK
PI3K
Shc
PLC
Rac/Cdc-GAP
Rac/CD C-GEF
PTEN
PtdIns3P
Grb
1,4,5-Insp3
1,2-Diacetin
Rac/CDC42
PAK
PDPK1
RAS-GRF
SOS
inositol 1,4,5-triphosphate
ITPK
PKC
PP2A
Ras
CALCIUM
PP5/2CB
MAP3K
AKT
Raf
calmodulin
NIK
CAMKK
MKK4/7
MKK3/6
DUSP
MKP
MEK1/2
IKBK
CAMK
JNK
p38
ERK1/3
NFKBI
calcineurin
Egr/Krox
RPS6K
MAPKAPK
PLA2
NF-kB
NF-AT
TNF family
Cytokine
CREB
ETS
SP
vesicular secretory pathway (Golgi -> membrane})
Prostaglandins, Thromboxane synthesis
Leukotriene synthesis
ARIADNE

Fig. 5 Molecular pathways triggered through FcεRI/mast cell axis. Representation of specific inflammatory and immunoregulatory pathways among the differentially expressed genes triggered through FcεRI/mast cell axis was analyzed using Pathway Studio® software (Ariadne Genomics, Rockville, MD) version 6.0. The software uses information available in the current literature to identify common pathways, targets, or regulators that are associated with the altered genes to generate biological interaction networks. Genes were linked to each other based on the published literature

(Ariadne Genomics, Rockville, MD) version 6.0. The software uses information available in the current literature to identify common pathways, targets, or regulators that are associated with the altered genes to generate biological interaction networks. Microarray expression data was imported into Pathway Studio® to graphically represent all known relationships and potential interactions between the differentially expressed genes. Pathway Studio software was used to identify a possible gene network that is differentially regulated in the mast cell-mediated classical pathway. Genes were linked to each other based on the published literature (Fig. 5).

Concluding remarks

Allergy and anaphylaxis comprise a wide spectrum of pathologies associated with the inappropriate activation of the immune system by environmental antigens [125]. Allergic responses to foods, insect bites, oral and injected medications, and other agents, remain huge problems, and are constantly increasing in society [125, 126]. Importantly, the FcεRI/mast cell axis is central to these immune reactions, and provides an attractive target for the inhibition of all IgE-mediated allergic diseases. Moreover, the modulation of this central axis has long been considered as a therapeutic strategy for various allergic disorders [50, 127]. Interestingly, clinical studies of allergic individuals using anti-IgE monoclonal antibody therapy have shown that the exploitation of this key axis is an effective approach to disease treatment [128, 129].

In addition, mast cells play a major role in the pathogenesis of inflammatory diseases such as asthma, atopic dermatitis, psoriasis, and interstitial cystitis, as well as a minor role in irritable bowel syndrome (IBS), inflammatory bowel disease (IBD), rheumatoid arthritis (RA), coronary artery disease (CAD), obesity, multiple sclerosis, and cancer [3, 4, 130]. As a result, the critical dissection of differentially expressed genes triggered through the FcεRI/mast cell axis and extensive exploration of their regulatory pathways through advancing high-throughput technologies, may allow us to selectively regulate these processes and help in the development of therapeutic modalities to potentially control and manage the exaggerated immune/inflammatory responses in allergy and anaphylaxis in the near future.

Acknowledgments The authors would like to thank the various research groups around the globe for their exceptional contributions in the basic, translational, and clinical aspects of allergy and anaphylaxis. We really regret the omission of any of these findings or contributions on the FcεRI/Mast cell axis in this review, which is mainly due to space limitations. We extend our immense appreciation to the graphics team of Beacon Biosoft (www.beaconbiosoft.com) for the fantastic figures in our review.

References

1. MacDonald SM, Rafnar T, Langdon J, Lichtenstein LM (1995) Molecular identification of an IgE-dependent histamine-releasing factor. Science 269(5224):688–690
2. Portier P, Richet C (1902) De l'action anaphylatique de certain venins. C R Soc Biol 54:170
3. Theoharides TC, Kalogeromitros D (2006) The critical role of mast cells in allergy and inflammation. Ann NY Acad Sci 1088:78–99
4. Theoharides TC, Alysandratos KD, Angelidou A, Delivanis DA, Sismanopoulos N, Zhang B, Asadi S, Vasiadi M, Weng Z, Miniati A, Kalogeromitros D (2010) Mast cells and inflammation. Biochim Biophys Acta. doi:10.1016/j.bbadis.2010.12.014
5. Metcalfe DD, Baram D, Mekori YA (1997) Mast cells. Physiol Rev 77(4):1033–1079
6. Munitz A, Piliponsky AM, Levi-Schaffer F (2003) IgE-independent activation of human mast cells indicates their role in the late-phase reaction of allergic inflammation. Cell Tissue Bank 4(1):25–28
7. Dudler T, Machado DC, Kolbe L, Annand RR, Rhodes N, Gelb MH, Koelsch K, Suter M, Helm BA (1995) A link between catalytic activity, IgE-independent mast cell activation, and allergenicity of bee venom phospholipase A2. J Immunol 155(5):2605–2613
8. Sullivan TJ, Hart DA, Streilein JW (1981) IgE-dependent release of inflammatory mediators from hamster mast cells in vitro. Adv Exp Med Biol 134:33–41
9. Heller RA, Schena M, Chai A, Shalon D, Bedilion T, Gilmore J, Woolley DE, Davis RW (1997) Discovery and analysis of inflammatory disease-related genes using cDNA microarrays. Proc Natl Acad Sci USA 94(6):2150–2155
10. Firoved AM, Ornatowski W, Deretic V (2004) Microarray analysis reveals induction of lipoprotein genes in mucoid pseudomonas aeruginosa: implications for inflammation in cystic fibrosis. Infect Immun 72(9):5012–5018
11. Spears R, Oakes R, Moore C, Bellinger LL, Hutchins B (2003) A determination of tumor necrosis factor expression in TMJ inflammation with the use of microarray analysis. J Dent Res 82(10):807–813
12. Olman MA, White KE, Ware LB, Cross MT, Zhu S, Matthay MA (2002) Microarray analysis indicates that pulmonary edema fluid from patients with acute lung injury mediates inflammation, mitogen gene expression, and fibroblast proliferation through bioactive interleukin-1. Chest 121((3 Suppl)):69S–70S
13. Gu J, Huang F, Yu D (2001) Analysis of inflammation related gene expression spectrum in ankylosing spondylitis patients using cDNA microarray. Zhonghua Yi Xue Za Zhi 81(17):1030–1034
14. Daeron M (1997) Fc receptor biology. Annu Rev Immunol 15:203–234
15. Unkeless JC, Scigliano E, Freedman VH (1988) Structure and function of human and murine receptors for IgG. Annu Rev Immunol 6:251–281
16. Ravetch JV, Kinet JP (1991) Fc receptors. Annu Rev Immunol 9:457–492
17. Vitetta ES, Uhr JW (1975) Immunoglobulin-receptors revisited. Science 189(4207):964–969
18. Davie JM (1975) Immunoglobulin receptors in lymphocytes. Adv Pathobiol (1):70–79
19. Keegan AD, Conrad DH (1990) The receptor for the Fc region of IgE. Springer Semin Immunopathol 12(4):303–326
20. Gould HJ, Sutton BJ, Beavil AJ, Beavil RL, McCloskey N, Coker HA, Fear D, Smurthwaite L (2003) The biology of IGE and the basis of allergic disease. Annu Rev Immunol 21:579–628
21. Rosenwasser LJ, Meng J (2005) Anti-CD23. Clin Rev Allergy Immunol 29(1):61–72
22. MacGlashan D Jr (2008) IgE receptor and signal transduction in mast cells and basophils. Curr Opin Immunol 20(6):717–723
23. Gilfillan AM, Rivera J (2009) The tyrosine kinase network regulating mast cell activation. Immunol Rev 228(1):149–169
24. Rivera J, Fierro NA, Olivera A, Suzuki R (2008) New insights on mast cell activation via the high affinity receptor for IgE. Adv Immunol 98:85–120
25. Kinet JP (1999) The high-affinity IgE receptor (Fc epsilon RI): from physiology to pathology. Annu Rev Immunol 17:931–972
26. Stingl G, Maurer D (1997) IgE-mediated allergen presentation via Fc epsilon RI on antigen-presenting cells. Int Arch Allergy Immunol 113(1–3):24–29
27. Uehara M, Sawai T (1989) Familial background of respiratory atopy. A factor of type I allergy to house dust mite in patients with atopic dermatitis. Arch Dermatol 125 (7):939–943
28. Smith SJ, Ying S, Meng Q, Sullivan MH, Barkans J, Kon OM, Sihra B, Larche M, Levi-Schaffer F, Kay AB (2000) Blood eosinophils from atopic donors express messenger RNA for the alpha, beta, and gamma subunits of the high-affinity IgE receptor (Fc epsilon RI) and intracellular, but not cell surface,

alpha subunit protein. J Allergy Clin Immunol 105 (2 Pt 1):309–317

29. Conrad DH, Campbell KA, Bartlett WC, Squire CM, Dierks SE (1994) Structure and function of the low affinity IgE receptor. Adv Exp Med Biol 347:17–30
30. Donnadieu E, Jouvin MH, Kinet JP (2000) A second amplifier function for the allergy-associated Fc (epsilon) RI-beta subunit. Immunity 12(5):515–523
31. Nadler MJ, Matthews SA, Turner H, Kinet JP (2000) Signal transduction by the high-affinity immunoglobulin E receptor Fc epsilon RI: coupling form to function. Adv Immunol 76:325–355
32. Kraft S, Bieber T (2001) Fc epsilon ri-mediated activation of transcription factors in antigen-presenting cells. Int Arch Allergy Immunol 125(1):9–15
33. Miller L, Blank U, Metzger H, Kinet JP (1989) Expression of high-affinity binding of human immunoglobulin E by transfected cells. Science 244(4902):334–337
34. Kawakami T, Galli SJ (2002) Regulation of mast-cell and basophil function and survival by IgE. Nat Rev Immunol 2(10):773–786
35. Jonsson F, Mancardi DA, Kita Y, Karasuyama H, Iannascoli B, Van Rooijen N, Shimizu T, Daeron M, Bruhns P (2011) Mouse and human neutrophils induce anaphylaxis. J Clin Invest 121(4):1484–1496
36. Reynolds DS, Serafin WE, Faller DV, Wall DA, Abbas AK, Dvorak AM, Austen KF, Stevens RL (1988) Immortalization of murine connective tissue-type mast cells at multiple stages of their differentiation by coculture of splenocytes with fibroblasts that produce Kirsten sarcoma virus. J Biol Chem 263(25):12783–12791
37. Xu D, Jiang HR, Li Y, Pushparaj PN, Kurowska-Stolarska M, Leung BP, Mu R, Tay HK, McKenzie AN, McInnes IB, Melendez AJ, Liew FY (2010) IL-33 exacerbates autoantibody-induced arthritis. J Immunol 184(5):2620–2626
38. Eady RA (1976) The mast cells: distribution and morphology. Clin Exp Dermatol 1(4):313–321
39. Metcalfe DD (2008) Mast cells and mastocytosis. Blood 112(4):946–956
40. Liew FY, Pitman NI, McInnes IB (2010) Disease-associated functions of IL-33: the new kid in the IL-1 family. Nat Rev Immunol 10(2):103–110
41. Kopec A, Panaszek B, Fal AM (2006) Intracellular signaling pathways in IgE-dependent mast cell activation. Arch Immunol Ther Exp (Warsz) 54(6):393–401
42. Kalesnikoff J, Huber M, Lam V, Damen JE, Zhang J, Siraganian RP, Krystal G (2001) Monomeric IgE stimulates signaling pathways in mast cells that lead to cytokine production and cell survival. Immunity 14(6):801–811
43. Baumruker T, Prieschl EE (2000) The role of sphingosine kinase in the signaling initiated at the high-affinity receptor for IgE (FcepsilonRI) in mast cells. Int Arch Allergy Immunol 122(2):85–90
44. Beaven MA, Metzger H (1993) Signal transduction by Fc receptors: the Fc epsilon RI case. Immunol Today 14(5):222–226
45. Barnes PJ, Chung KF, Page CP (1998) Inflammatory mediators of asthma: an update. Pharmacol Rev 50(4):515–596
46. Gilfillan AM, Tkaczyk C (2006) Integrated signalling pathways for mast-cell activation. Nat Rev Immunol 6(3):218–230
47. Kambayashi T, Koretzky GA (2007) Proximal signaling events in Fc epsilon RI-mediated mast cell activation. J Allergy Clin Immunol 119 (3):544–552; (quiz 553-544)
48. Kitaura J, Eto K, Kinoshita T, Kawakami Y, Leitges M, Lowell CA, Kawakami T (2005) Regulation of highly cytokinergic IgE-induced mast cell adhesion by Src, Syk, Tec, and protein kinase C family kinases. J Immunol 174(8):4495–4504
49. Rivera J (2002) Molecular adapters in Fc(epsilon)RI signaling and the allergic response. Curr Opin Immunol 14(6):688–693
50. Turner H, Kinet JP (1999) Signalling through the high-affinity IgE receptor Fc epsilonRI. Nature 402(6760 Suppl):B24–B30
51. Goldstein B, Faeder JR, Hlavacek WS, Blinov ML, Redondo A, Wofsy C (2002) Modeling the early signaling events mediated by FcepsilonRI. Mol Immunol 38(16–18):1213–1219
52. Craig AW, Zirngibl R, Williams K, Cole LA, Greer PA (2001) Mice devoid of fer protein-tyrosine kinase activity are viable and fertile but display reduced cortactin phosphorylation. Mol Cell Biol 21(2):603–613
53. Choi OH, Kim JH, Kinet JP (1996) Calcium mobilization via sphingosine kinase in signalling by the Fc epsilon RI antigen receptor. Nature 380(6575):634–636
54. Aarthi JJ, Darendeliler MA, Pushparaj PN (2011) Dissecting the role of S1P/S1PR axis in health and disease. J Dent Res 90(7):841–854
55. Jolly PS, Bektas M, Olivera A, Gonzalez-Espinosa C, Proia RL, Rivera J, Milstien S, Spiegel S (2004) Transactivation of sphingosine-1-phosphate receptors by FcepsilonRI triggering is required for normal mast cell degranulation and chemotaxis. J Exp Med 199(7):959–970
56. Urtz N, Olivera A, Bofill-Cardona E, Csonga R, Billich A, Mechtcheriakova D, Bornancin F, Woisetschlager M, Rivera J, Baumruker T (2004) Early activation of sphingosine kinase in mast cells and recruitment to FcepsilonRI are mediated by its interaction with Lyn kinase. Mol Cell Biol 24(19):8765–8777
57. Malveaux FJ, Conroy MC, Adkinson NF, Jr., Lichtenstein LM (1978) IgE receptors on human basophils. Relationship to serum IgE concentration. J Clin Invest 62 (1):176–181
58. Hsu C, MacGlashan D Jr (1996) IgE antibody up-regulates high affinity IgE binding on murine bone marrow-derived mast cells. Immunol Lett 52(2–3):129–134
59. Lantz CS, Yamaguchi M, Oettgen HC, Katona IM, Miyajima I, Kinet JP, Galli SJ (1997) IgE regulates mouse basophil Fc epsilon RI expression in vivo. J Immunol 158(6):2517–2521
60. Yamaguchi M, Sayama K, Yano K, Lantz CS, Noben-Trauth N, Ra C, Costa JJ, Galli SJ (1999) IgE enhances Fc epsilon receptor I expression and IgE-dependent release of histamine and lipid mediators from human umbilical cord blood-derived mast cells: synergistic effect of IL-4 and IgE on human mast cell Fc epsilon receptor I expression and mediator release. J Immunol 162(9):5455–5465
61. Woolhiser MR, Okayama Y, Gilfillan AM, Metcalfe DD (2001) IgG-dependent activation of human mast cells following up-regulation of FcgammaRI by IFN-gamma. Eur J Immunol 31(11):3298–3307
62. Gurish MF, Bryce PJ, Tao H, Kisselgof AB, Thornton EM, Miller HR, Friend DS, Oettgen HC (2004) IgE enhances parasite clearance and regulates mast cell responses in mice infected with *Trichinella spiralis*. J Immunol 172(2):1139–1145
63. Oettgen HC, Geha RS (2001) IgE regulation and roles in asthma pathogenesis. J Allergy Clin Immunol 107(3):429–440
64. Oettgen HC, Martin TR, Wynshaw-Boris A, Deng C, Drazen JM, Leder P (1994) Active anaphylaxis in IgE-deficient mice. Nature 370(6488):367–370
65. Pushparaj PN, Tay HK, H’Ng SC, Pitman N, Xu D, McKenzie A, Liew FY, Melendez AJ (2009) The cytokine interleukin-33 mediates anaphylactic shock. Proc Natl Acad Sci USA 106(24):9773–9778
66. Maurer D, Fiebiger E, Reininger B, Wolff-Winiski B, Jouvin MH, Kilgus O, Kinet JP, Stingl G (1994) Expression of functional high affinity immunoglobulin E receptors (Fc epsilon RI) on monocytes of atopic individuals. J Exp Med 179(2):745–750
67. Reischl IG, Dubois GR, Peiritsch S, Brown KS, Wheat L, Woisetschlager M, Mudde GC (2000) Regulation of Fc

epsilonRI expression on human monocytic cells by ligand and IL-4. Clin Exp Allergy 30(7):1033–1040

68. Cookson WO, Sharp PA, Faux JA, Hopkin JM (1989) Linkage between immunoglobulin E responses underlying asthma and rhinitis and chromosome 11q. Lancet 1(8650):1292–1295
69. Nishiyama C, Akizawa Y, Nishiyama M, Tokura T, Kawada H, Mitsuishi K, Hasegawa M, Ito T, Nakano N, Okamoto A, Takagi A, Yagita H, Okumura K, Ogawa H (2004) Polymorphisms in the Fc epsilon RI beta promoter region affecting transcription activity: a possible promoter-dependent mechanism for association between Fc epsilon RI beta and atopy. J Immunol 173(10):6458–6464
70. MacGlashan D, Jr., Lichtenstein LM, McKenzie-White J, Chichester K, Henry AJ, Sutton BJ, Gould HJ (1999) Upregulation of FcepsilonRI on human basophils by IgE antibody is mediated by interaction of IgE with FcepsilonRI. J Allergy Clin Immunol 104 (2 Pt 1):492–498
71. MacGlashan D Jr, McKenzie-White J, Chichester K, Bochner BS, Davis FM, Schroeder JT, Lichtenstein LM (1998) In vitro regulation of Fc epsilon RIalpha expression on human basophils by IgE antibody. Blood 91(5):1633–1643
72. Borkowski TA, Jouvin MH, Lin SY, Kinet JP (2001) Minimal requirements for IgE-mediated regulation of surface Fc epsilon RI. J Immunol 167(3):1290–1296
73. Toru H, Ra C, Nonoyama S, Suzuki K, Yata J, Nakahata T (1996) Induction of the high-affinity IgE receptor (Fc epsilon RI) on human mast cells by IL-4. Int Immunol 8(9):1367–1373
74. Capron A, Dombrowicz D, Capron M (1999) Regulation of the immune response in experimental and human schistosomiasis: the limits of an attractive paradigm. Microbes Infect 1(7):485–490
75. Simon MR, Cooper KD, Norris RB, Blok B, King CL (1995) Antigen presenting cell-independent cytokine and spontaneous in vitro IgE production in patients with atopic dermatitis: increased interferon-gamma production and lack of effects of in vivo low-dose interferon-gamma treatment. J Allergy Clin Immunol 96(1):84–91
76. Fouilloux I, Duplan MB, Baroukh B, Cherruau M, Saffar JL, Lesclous P (2006) Mast cell activation and degranulation occur early during induction of periosteal bone resorption. Bone 38(1):59–66
77. Gurish MF, Ghildyal N, Arm J, Austen KF, Avraham S, Reynolds D, Stevens RL (1991) Cytokine mRNA are preferentially increased relative to secretory granule protein mRNA in mouse bone marrow-derived mast cells that have undergone IgE-mediated activation and degranulation. J Immunol 146(5):1527–1533
78. Morrison DC, Roser JF, Cochrane CG, Henson PM (1975) The initiation of mast cell degranulation: activation at the cell membrane. J Immunol 114(3):966–970
79. Kaliner M, Shelhamer JH, Ottesen EA (1982) Effects of infused histamine: correlation of plasma histamine levels and symptoms. J Allergy Clin Immunol 69(3):283–289
80. Morel DR, Skoskiewicz M, Robinson DR, Bloch KJ, Hoaglin DC, Zapol WM (1991) Leukotrienes, thromboxane A2, and prostaglandins during systemic anaphylaxis in sheep. Am J Physiol 261 (3 Pt 2):H782–792
81. Vadas P, Gold M, Perelman B, Liss GM, Lack G, Blyth T, Simons FE, Simons KJ, Cass D, Yeung J (2008) Platelet-activating factor, PAF acetylhydrolase, and severe anaphylaxis. N Engl J Med 358(1):28–35
82. Pretolani M, Lellouch-Tubiana A, Lefort J, Bachelet M, Vargaftig BB (1989) PAF-acether and experimental anaphylaxis as a model for asthma. Int Arch Allergy Appl Immunol 88(1–2):149–153
83. Felix SB, Baumann G, Hashemi T, Niemczyk M, Ahmad Z, Berdel WE (1990) Characterization of cardiovascular events mediated by platelet activating factor during systemic anaphylaxis. J Cardiovasc Pharmacol 15(6):987–997
84. Gordon JR, Galli SJ (1991) Release of both preformed and newly synthesized tumor necrosis factor alpha (TNF-alpha)/cachectin by mouse mast cells stimulated via the Fc epsilon RI. A mechanism for the sustained action of mast cell-derived TNF-alpha during IgE-dependent biological responses. J Exp Med 174 (1):103–107
85. Comaish JS (1965) The role of the mast cells and basophils in hypersensitivity. Br J Dermatol 77:92–97
86. von Bubnoff D, Novak N, Kraft S, Bieber T (2003) The central role of FcepsilonRI in allergy. Clin Exp Dermatol 28(2):184–187
87. Maurer D, Ebner C, Reininger B, Petzelbauer P, Fiebiger E, Stingl G (1997) Mechanisms of Fc epsilon RI-IgE-facilitated allergen presentation by dendritic cells. Adv Exp Med Biol 417:175–178
88. Kalinski P, Hilkens CM, Wierenga EA, Kapsenberg ML (1999) T-cell priming by type-1 and type-2 polarized dendritic cells: the concept of a third signal. Immunol Today 20(12):561–567
89. Karras JG, McGraw K, McKay RA, Cooper SR, Lerner D, Lu T, Walker C, Dean NM, Monia BP (2000) Inhibition of antigen-induced eosinophilia and late phase airway hyperresponsiveness by an IL-5 antisense oligonucleotide in mouse models of asthma. J Immunol 164(10):5409–5415
90. Kay AB (2001) Allergy and allergic diseases. First of two parts. N Engl J Med 344 (1):30-37
91. Berin MC, Mayer L (2009) Immunophysiology of experimental food allergy. Mucosal Immunol 2(1):24–32
92. Muller UR (2001) New developments in the diagnosis and treatment of hymenoptera venom allergy. Int Arch Allergy Immunol 124(4):447–453
93. Ying S, Robinson DS, Meng Q, Barata LT, McEuen AR, Buckley MG, Walls AF, Askenase PW, Kay AB (1999) C-C chemokines in allergen-induced late-phase cutaneous responses in atopic subjects: association of eotaxin with early 6-hour eosinophils, and of eotaxin-2 and monocyte chemoattractant protein-4 with the later 24-hour tissue eosinophilia, and relationship to basophils and other C-C chemokines (monocyte chemoattractant protein-3 and RANTES). J Immunol 163(7):3976–3984
94. Robinson DS, Hamid Q, Jacobson M, Ying S, Kay AB, Durham SR (1993) Evidence for Th2-type T helper cell control of allergic disease in vivo. Springer Semin Immunopathol 15(1):17–27
95. Chang-Yeung M, Lam S, Kennedy SM, Frew AJ (1994) Persistent asthma after repeated exposure to high concentrations of gases in pulpmills. Am J Respir Crit Care Med 149(6):1676–1680
96. Arguelles M, Blanco I (1983) Inflammatory bronchial polyps associated with asthma. Arch Intern Med 143(3):570–571
97. Macfarlane AJ, Kon OM, Smith SJ, Zeibecoglou K, Khan LN, Barata LT, McEuen AR, Buckley MG, Walls AF, Meng Q, Humbert M, Barnes NC, Robinson DS, Ying S, Kay AB (2000) Basophils, eosinophils, and mast cells in atopic and nonatopic asthma and in late-phase allergic reactions in the lung and skin. J Allergy Clin Immunol 105 (1 Pt 1):99–107
98. Elmgreen J, Stahl Skov P, Permin H, Binder V, Heugh Wandall J, Norn S (1984) Type I allergy to normal cellular constituents in chronic inflammatory bowel disease? Results from basophil histamine release test compared with total IgE and antinuclear antibodies. Allergy 39(1):23–28
99. Steinhoff M, Vergnolle N, Young SH, Tognetto M, Amadesi S, Ennes HS, Trevisani M, Hollenberg MD, Wallace JL, Caughey GH, Mitchell SE, Williams LM, Geppetti P, Mayer EA, Bunnett

NW (2000) Agonists of proteinase-activated receptor 2 induce inflammation by a neurogenic mechanism. Nat Med 6(2):151–158
100. Holt PG, Macaubas C, Stumbles PA, Sly PD (1999) The role of allergy in the development of asthma. Nature 402((6760 Suppl)):B12–B17
101. Dolovich J, Hargreave FE, Chalmers R, Shier KJ, Gauldie J, Bienenstock J (1973) Late cutaneous allergic responses in isolated IgE-dependent reactions. J Allergy Clin Immunol 52(1):38–46
102. Milgrom H (2002) Attainments in atop: special aspects of allergy and IgE. Adv Pediatr 49:273–297
103. Gospos A, Dreikhausen U, Dartsch DC, Szamel M, Hockertz S (2001) Development of an allergy test model: activation of human mast cells with potentially allergenic substances. Toxicology 166(1–2):91–96
104. van Halteren AG, van der Cammen MJ, Biewenga J, Savelkoul HF, Kraal G (1997) IgE and mast cell response on intestinal allergen exposure: a murine model to study the onset of food allergy. J Allergy Clin Immunol 99 (1 Pt 1):94–99
105. Toru H, Pawankar R, Ra C, Yata J, Nakahata T (1998) Human mast cells produce IL-13 by high-affinity IgE receptor cross-linking: enhanced IL-13 production by IL-4-primed human mast cells. J Allergy Clin Immunol 102(3):491–502
106. Haselden BM, Kay AB, Larche M (1999) Immunoglobulin E-independent major histocompatibility complex-restricted T cell peptide epitope-induced late asthmatic reactions. J Exp Med 189(12):1885–1894
107. Herz U, Botchkarev VA, Paus R, Renz H (2004) Increased airway responsiveness, allergy-type-I skin responses and systemic anaphylaxis in a humanized-severe combined immunodeficiency mouse model. Clin Exp Allergy 34(3):478–487
108. Berger P, N'Guyen C, Buckley M, Scotto-Gomez E, Marthan R, Tunon-de-Lara JM (2002) Passive sensitization of human airways induces mast cell degranulation and release of tryptase. Allergy 57(7):592–599
109. Fahy JV, Fleming HE, Wong HH, Liu JT, Su JQ, Reimann J, Fick RB Jr, Boushey HA (1997) The effect of an anti-IgE monoclonal antibody on the early- and late-phase responses to allergen inhalation in asthmatic subjects. Am J Respir Crit Care Med 155(6):1828–1834
110. Benoist C, Mathis D (2002) Mast cells in autoimmune disease. Nature 420(6917):875–878
111. Pachiappan A, Thwin MM, Manikandan J, Gopalakrishnakone P (2005) Glial inflammation and neurodegeneration induced by candoxin, a novel neurotoxin from *Bungarus candidus* venom: global gene expression analysis using microarray. Toxicon 46(8):883–899
112. Shen F, Ruddy MJ, Plamondon P, Gaffen SL (2005) Cytokines link osteoblasts and inflammation: microarray analysis of interleukin-17- and TNF-alpha-induced genes in bone cells. J Leukoc Biol 77(3):388–399
113. Rundle CH, Wang H, Yu H, Chadwick RB, Davis EI, Wergedal JE, Lau KH, Mohan S, Ryaby JT, Baylink DJ (2006) Microarray analysis of gene expression during the inflammation and endochondral bone formation stages of rat femur fracture repair. Bone 38(4):521–529
114. Shiue HS, Lee YS, Tsai CN, Hsueh YM, Sheu JR, Chang HH (2008) DNA microarray analysis of the effect on inflammation in patients treated with acupuncture for allergic rhinitis. J Altern Complement Med 14(6):689–698
115. Jayapal M, Tay HK, Reghunathan R, Zhi L, Chow KK, Rauff M, Melendez AJ (2006) Genome-wide gene expression profiling of human mast cells stimulated by IgE or FcepsilonRI-aggregation reveals a complex network of genes involved in inflammatory responses. BMC Genomics 7:210
116. Saito H, Nakajima T, Matsumoto K (2001) Human mast cell transcriptome project. Int Arch Allergy Immunol 125(1):1–8
117. Kuramasu A, Kubota Y, Matsumoto K, Nakajima T, Sun XM, Watanabe T, Saito H, Ohtsu H (2001) Identification of novel mast cell genes by serial analysis of gene expression in cord blood-derived mast cells. FEBS Lett 498(1):37–41
118. Wakahara S, Fujii Y, Nakao T, Tsuritani K, Hara T, Saito H, Ra C (2001) Gene expression profiles for Fc epsilon RI, cytokines and chemokines upon Fc epsilon RI activation in human cultured mast cells derived from peripheral blood. Cytokine 16(4):143–152
119. Nakajima T, Matsumoto K, Suto H, Tanaka K, Ebisawa M, Tomita H, Yuki K, Katsunuma T, Akasawa A, Hashida R, Sugita Y, Ogawa H, Ra C, Saito H (2001) Gene expression screening of human mast cells and eosinophils using high-density oligonucleotide probe arrays: abundant expression of major basic protein in mast cells. Blood 98(4):1127–1134
120. Nakamura R, Ishida S, Ozawa S, Saito Y, Okunuki H, Teshima R, Sawada J (2002) Gene expression profiling of Ca^{2+} -ATPase inhibitor DTBHQ and antigen-stimulated RBL-2H3 mast cells. Inflamm Res 51(12):611–618
121. Nakajima T, Inagaki N, Tanaka H, Tanaka A, Yoshikawa M, Tamari M, Hasegawa K, Matsumoto K, Tachimoto H, Ebisawa M, Tsujimoto G, Matsuda H, Nagai H, Saito H (2002) Marked increase in CC chemokine gene expression in both human and mouse mast cell transcriptomes following Fcepsilon receptor I cross-linking: an interspecies comparison. Blood 100(12):3861–3868
122. Okumura S, Kashiwakura J, Tomita H, Matsumoto K, Nakajima T, Saito H, Okayama Y (2003) Identification of specific gene expression profiles in human mast cells mediated by Toll-like receptor 4 and FcepsilonRI. Blood 102(7):2547–2554
123. Babina M, Guhl S, Starke A, Kirchhof L, Zuberbier T, Henz BM (2004) Comparative cytokine profile of human skin mast cells from two compartments–strong resemblance with monocytes at baseline but induction of IL-5 by IL-4 priming. J Leukoc Biol 75(2):244–252
124. Matsuda K, Piliponsky AM, Iikura M, Nakae S, Wang EW, Dutta SM, Kawakami T, Tsai M, Galli SJ (2005) Monomeric IgE enhances human mast cell chemokine production: IL-4 augments and dexamethasone suppresses the response. J Allergy Clin Immunol 116(6):1357–1363
125. Simons FE (2010) World Allergy Organization survey on global availability of essentials for the assessment and management of anaphylaxis by allergy-immunology specialists in health care settings. Ann Allergy Asthma Immunol 104(5):405–412
126. Demain JG, Minaei AA, Tracy JM (2010) Anaphylaxis and insect allergy. Curr Opin Allergy Clin Immunol 10(4):318–322
127. Garman SC, Sechi S, Kinet JP, Jardetzky TS (2001) The analysis of the human high affinity IgE receptor Fc epsilon Ri alpha from multiple crystal forms. J Mol Biol 311(5):1049–1062
128. Leavy O Therapeutic antibodies: past, present and future. Nat Rev Immunol 10 (5):297
129. Chan AC, Carter PJ Therapeutic antibodies for autoimmunity and inflammation. Nat Rev Immunol 10 (5):301–316
130. Bhattacharjee RN, Banerjee B, Akira S, Hande MP (2010) Telomere-mediated chromosomal instability triggers TLR4 induced inflammation and death in mice. PLoS One 5(7):e11873

ws.org. Guest (guest) IP: 24.117.150.43 On: Tue, 14 Oct 2025 03:05:08

ANNUAL REVIEWS

Annual Review of Genomics and Human Genetics

The Genetics and Genomics of Asthma

Saffron A.G. Willis-Owen, William O.C. Cookson, and Miriam F. Moffatt

National Heart and Lung Institute, Imperial College London, London SW7 2AZ, United Kingdom; email: s.willis-owen@imperial.ac.uk, w.cookson@imperial.ac.uk, m.moffatt@imperial.ac.uk

Annu. Rev. Genom. Hum. Genet. 2018. 19:223–46

The *Annual Review of Genomics and Human Genetics* is online at genom.annualreviews.org

https://doi.org/10.1146/annurev-genom-083117-021651

Copyright © 2018 by Annual Reviews.
All rights reserved

ANNUAL REVIEWS CONNECT
www.annualreviews.org
- Download figures
- Navigate cited references
- Keyword search
- Explore related articles
- Share via email or social media

Keywords

asthma, genetics, microbiome, epigenetics, infection, co-morbidity

Abstract

Asthma is a common, clinically heterogeneous disease with strong evidence of heritability. Progress in defining the genetic underpinnings of asthma, however, has been slow and hampered by issues of inconsistency. Recent advances in the tools available for analysis—assaying transcription, sequence variation, and epigenetic marks on a genome-wide scale—have substantially altered this landscape. Applications of such approaches are consistent with heterogeneity at the level of causation and specify patterns of commonality with a wide range of alternative disease traits. Looking beyond the individual as the unit of study, advances in technology have also fostered comprehensive analysis of the human microbiome and its varied roles in health and disease. In this article, we consider the implications of these technological advances for our current understanding of the genetics and genomics of asthma.

Downloaded from www.annualreviews.org. Guest (guest) IP: 24.117.150.43 On: Tue, 14 Oct 2025 03:05:08

INTRODUCTION

Asthma is a chronic disease of the airways defined by its symptoms, which include reversible airflow obstruction, inflammation, and bronchial hyperresponsiveness. The ancient Egyptians, Greeks, and Romans made reference to the symptoms of asthma, and today the disease is estimated to affect 235–334 million people worldwide (44, 53).

Asthma is heterogeneous in both its clinical presentation and its prognosis. The International Classification of Diseases diagnostic system primarily groups asthma patients according to the severity and persistence of symptoms and the presence of complications (152). Statistical approaches to subtype asthma likewise observe substructure among symptoms (such as atopic status, wheeze, and cough) (67), but the nature of this structure may depend on choices relating to variable inclusion and methodology (34), limiting the overall clinical application of such phenotype-driven techniques.

Asthma has a substantial genetic component, with estimates of heritability ranging from 35% to 95% (116). As such, the genomic era brought with it hopes of molecular subtyping and associated personalization of treatment plans for the disease, as has been seen in the case of breast cancer (20, 28). However, given that the heritability of asthma is incomplete (<1), environmental factors must also account for a fraction of the phenotypic variance. Consistent with this, epidemiological studies have identified a range of pre-, peri-, and postnatal environmental factors, including mode of delivery, diet, and early lower respiratory tract infection, that confer relative risk or protection. Attempts to map the genetic architecture of asthma have identified a broad spectrum of potential contributory genes. Many of these genes demonstrate inconsistent patterns of replication between cohorts, most likely reflecting a combination of true positive and true negative results and the unseen hand of gene × environment and gene × gene interaction. Nevertheless, a handful of genes have shown robust association with asthma. While these effects individually account for only a minority of phenotypic variance, they provide novel insights into the underlying pathogenic processes as well as potential sources of interaction with the environment. Here, we review the genetic and genomic factors that are currently implicated in the pathobiology of asthma and consider the ability of these factors to differentiate between patient groups.

GENETIC LINKAGE AND POSITIONAL CLONING

Prior to the advent of high-throughput technologies, the cost of genotyping was substantial. As such, genome-wide, hypothesis-free approaches were limited to relatively small panels of informative (multiallelic) markers distributed approximately evenly across the genome. Genotype at these sites was determined in families with multiple affected individuals, allowing identification of contiguous regions that showed co-transmission with the trait of interest. Broad loci specified as harboring susceptibility variants could then be explored at a higher resolution through positional cloning or positional candidate gene studies, ideally identifying genes or the individual sequence variants responsible. This technique, highly successful in the arena of Mendelian traits, was the mainstay of early genetic studies of asthma, with more than 25 such studies reported between 1996 and 2006 (reviewed in 18, 35). While these studies led to the identification of only eight genes, *ADAM33*, *DPP10*, *PHF11*, *NPSR1*, *HLA-G*, *CYFIP2*, *IRAK3*, and *OPN3* (3, 11, 78, 112, 114, 142, 151, 162), that putatively contribute to the genetic etiology of the disease, each of small effect, all but one were novel, with no previously suspected role in asthma.

The eight positionally cloned putative susceptibility genes (listed in **Table 1**) are diverse in their described functions, ranging from cytokine and chemokine processing (*DPP10*) (3) to class-switch recombination (*PHF11*) and airway remodeling (*ADAM33*). *ADAM33* was the first gene

Downloaded from www.annualreviews.org. Guest (guest) IP: 24.117.150.43 On: Tue, 14 Oct 2025 03:05:08

Table 1 Positionally cloned putative asthma genes and their signals in genome-wide association studies

Positionally cloned gene	Region	Description	Mapped traits accompanied by low *p* values
ADAM33 (ADAM metallopeptidase domain 33) (142)	20p13	Metalloprotease	None
DPP10 (dipeptidyl peptidase–like 10) (3)	2q14	Serine protease	Severe influenza A (H1N1) infection (51), response to antidepressants in depression (83), survival in microsatellite instability low/stable colorectal cancer (154), gut microbiome composition (winter) (30), subjective well-being (117), systolic blood pressure in sickle cell anemia (15), bipolar disorder (71), obesity-related traits (25), response to paliperidone in schizophrenia (negative Marder score) (82), and suicide attempts in major depressive disorder (50)
PHF11 (PHD finger protein 11) (162)	13q14	Double-stranded DNA–binding protein	Breastfeeding duration (24) and cardiac hypertrophy (119)
NPSR1 (neuropeptide S receptor 1) (78)	7p14	G-protein-coupled receptor	IgG glycosylation (81), postbronchodilator FEV1 (94), and postbronchodilator FEV1/FVC ratio (94)
HLA-G (major histocompatibility complex, class I, G) (112)	6p21	Major histocompatibility complex antigen	Autism spectrum disorder or schizophrenia (10), eosinophil percentage of granulocytes (9), eosinophil percentage of white cells (9), neutrophil percentage of granulocytes (9), Crohn's disease (75), lung adenocarcinoma (103), IgE levels (57), and childhood ear infection (123)
CYFIP2 (cytoplasmic FMR1-interacting protein 2) (114)	5q33	G-protein modulator	Lung function (FEV1/FVC) (144) and IgG glycosylation (81)
IRAK3 (interleukin 1 receptor–associated kinase 3) (11)	12q14	Kinase	Smooth-surface caries (160)
OPN3 (opsin 3) (151)	1qter	G-protein-coupled receptor	Daytime sleep phenotypes (134)

This table summarizes the genes positionally cloned for asthma per se and includes their genomic location. The publicly available GWAS Catalog (19) was systematically queried for these genes, and traits accompanied by a p value $\leq 5 \times 10^{-6}$ at the time of writing are listed. Abbreviations: FEV1, forced expiratory volume measured during the first second of forced breath; FVC, forced vital capacity.

to be positionally cloned for asthma. Its transcripts show a preferential distribution in cells of mesenchymal origin (139, 142), and its sequence variants associate not only with asthma per se but also with bronchial hyperresponsiveness, early-life lung function, lung function decline across the life course, and chronic obstructive pulmonary disease (31), consistent with the gene's proposed role in airway remodeling.

Unfortunately, as with many other so-called complex (or multifactorial) traits, positionally cloned putative asthma susceptibility genes have been hampered by issues of inconsistent replication. While there are a host of competing explanations for this phenomenon, contributory factors may include false positives, allelic heterogeneity, natural variation in linkage disequilibrium between causative and genotyped sequence variants among assayed populations, the effects of other genetic variants and environments, and/or gene × gene or gene × environment interactions. By way of example, in 2003 *PHF11* was posited to be responsible for the chromosome 13q14 linkage peak that influences asthma and total serum immunoglobulin E (IgE) levels, as variants within this gene demonstrated the strongest evidence of association with the tested traits (162). Consistent with this hypothesis, *PHF11* was also shown to accelerate class-switch recombination to IgE in murine activated B cells (68). Although the chromosomal interval that contains *PHF11* has been implicated in asthma, atopy, and IgE traits in several linkage studies and the association between *PHF11* and asthma has been independently replicated, replication has not been successful across the full range of populations tested (101). Within the genome, *PHF11* is positioned in a linkage disequilibrium block that includes both *PHF11* and its immediate physical neighbor, *SETDB2* (SET domain bifurcated 2). These two genes, separated by an interval of just 4 kb, are transcribed in the same direction and form fusion transcripts in some tissues (162). A recent analysis of this region identified a functional variant located within the 5′ untranslated region of *SETDB2* that is capable of regulating the transcription of *SETDB2* and shows significant association with total serum IgE (66). Such situations are difficult to disentangle but highlight the potential for differing linkage disequilibrium patterns and/or allelic heterogeneity to manifest as inconsistent replication.

CANDIDATE GENES

The prohibitively high cost of genotyping combined with the paucity of available markers in the pre-genome era rendered high-resolution, hypothesis-free approaches impractical. As such, the vast majority of high-resolution studies were limited to individual genes whose candidacy was suggested by alternative, converging lines of evidence, including positional, mechanistic, and/or theoretical data. By design, this approach precluded identifying truly novel components of disease pathobiology, although it did instead provide a tool to ratify or refute existing candidates. The cost efficacy of the approach resulted in a large increase in the number of candidate gene association studies. However, these studies were invariably carried out in small, low-powered cohorts that are also prone to false positives, and consequently many of the results failed to replicate consistently between populations. This phenomenon is not limited to the context of asthma, but is widely recognized across a broad spectrum of complex traits (64, 69). For example, in 2002 a cross-trait analysis of candidate gene associations showed that less than 4% of associations studied three or more times were consistently replicated (64).

The design of genetic association studies is relatively simplistic. For categorical traits like disease status, allele frequencies at a small set of variants positioned in and around a gene of interest or the combinations thereof (haplotypes) are compared between cases and controls. These variants can take any format, such as insertion–deletion (indel) polymorphisms or single-nucleotide polymorphisms (SNPs). As of 2006, prior to publication of the first asthma genome-wide association study (GWAS), candidate gene studies had implicated more than 100 genes in the etiology of asthma (115). Of these, fewer than 10% had been replicated in more than 10 studies, even when the phenotype definition was expanded to include both asthma and atopy phenotypes and when the gene was considered the unit of replication, rather than individual variants or combinations thereof (115).

These relatively more promising asthma candidate genes included the prototypical type 2 cytokine genes *IL4* (interleukin 4) and *IL13* (interleukin 13) and the IL4 receptor gene *IL4RA* (now known as *IL4R*). Additionally, implicated in more than 10 studies was the beta-2 adrenergic receptor gene *ADRB2*, the candidacy of which has been supported by the efficacy of beta agonists in asthma and the gene's influence on beta-agonist responsiveness (100), and the positional candidate *ADAM33*. Other frequently replicated genes included the gene encoding the beta subunit of the high-affinity IgE receptor (*FCER1B*, now known as *MS4A2*), the proinflammatory cytokine gene *TNFA*, the surface antigen gene *CD14*, and the major histocompatibility locus genes *HLA-DRB1* and *HLA-DQB1*. That said, experimentally, some candidate genes have received considerably more attention than others, and factors such as publication bias, which influences the number of

Downloaded from www.annualreviews.org. Guest (guest) IP: 24.117.150.43 On: Tue, 14 Oct 2025 03:05:08

negative replications reported in the literature, further confound the direct interpretation of results. Unfortunately, large-scale meta-analyses capable of providing a clear synthesis of all available data are few and far between, not least owing to the diversity of the populations and environments studied. Furthermore, those that do exist often fail to detect a consistent overall effect, casting doubt on the credibility of many targets identified during the candidate gene era.

GENOME-WIDE ASSOCIATION STUDIES

GWASs leverage the high frequency of SNPs across the genome, as defined through sequencing, to enable the application of allelic association on a genome-wide scale. This technique is founded on the common disease–common variant principle, which predicts that the genetic architecture of common diseases is in large part composed of common, low-penetrance variants. Since GWASs are agnostic to what part(s) of the genome might be associated, they enable the detection of previously undescribed and unsuspected genetic components, assuming they are present in the population at an appreciable frequency. Variants detected as significantly associated in a GWAS are not necessarily themselves pathogenic, but may be in linkage disequilibrium with other, perhaps rarer, untyped variants that may be the source of the association.

The first GWAS of asthma was carried out in 2007 (109), and today, according to the publicly available GWAS Catalog (19), 38 loci, or genomic regions, have been associated with asthma per se at a p-value threshold of $p < 5 \times 10^{-8}$, the standard threshold commonly applied in GWASs (**Table 2**). This number markedly increases if intermediate phenotypes and disease subtypes are additionally considered. Among these 38 asthma-associated loci, the most consistently replicated, through GWAS and non-GWAS positional candidate gene studies and across a diverse range of ethnic backgrounds, is the cluster of genes on chromosome 17q12–21. This locus, originally brought to light in the first GWAS of asthma in 2007 (109), harbors sequence variants capable of regulating the expression level of several physically colocalized genes. These genes include *ORMDL3* (orosomucoid-like 3) (58, 90, 109), *GSDMB* (gasdermin B) (58), and *GSDMA* (gasdermin A) (60, 90). Variants that are capable of regulating the expression of local genes are referred to as *cis*-acting expression quantitative trait loci (eQTLs), and at least in the case of *ORMDL3*, they appear to have cell type–specific effects (89).

Large-scale studies powered to identify subgroup effects have further shown that the 17q12–21 signal is specific to childhood-onset disease. In addition, the relationship between genotype and disease appears to be moderated by early environmental exposures, including tobacco smoke (17), respiratory infection (130), and place of residence (132). The pathogenic mechanism or mechanisms that mediate these effects are as yet unclear. At least two of the genes positioned within the bounds of the asthma-associated 17q12–21 locus, however, appear to participate in similar processes underlying airway remodeling and responsiveness in asthma. Functional studies in mice show that *ORMDL3* expression affects airway remodeling (107), including airway smooth-muscle proliferation and contractility (21), with airway remodeling preceding airway inflammation. Similarly, transgenic mouse studies of *GSDMB* also suggest roles in airway remodeling and responsiveness but in the absence of lung inflammation (29). *ORMDL3*, an endoplasmic reticulum gene, mediates sphingolipid homeostasis and the unfolded protein (stress) response, while the function of *GSDMB* (also known as *GSDML*) is largely undescribed. In addition, there is growing evidence that expression of *ORMDL3* (or its mouse counterpart, *Ormdl3*) in the lung influences the host response to fungal (*Alternaria*) challenge (92) as well as the antiviral response to rhinovirus infection—showing a significant relationship with both viral load and the extent of airway inflammation (131). These data therefore suggest a mechanism for the interaction between genotype at the 17q12–21 locus and early respiratory infection in the development of asthma (130).

Downloaded from www.annualreviews.org. Guest (guest) IP: 24.117.150.43 On: Tue, 14 Oct 2025 03:05:08

Table 2 Loci implicated in the etiology of asthma via genome-wide association studies with a p value $< 5 \times 10^{-8}$

Region	Reported gene(s)	Reference(s)
1p36.22	*PEX14*	123
1q21.3	*IL6R*	45
1q23.1	*PYHIN1*	137
1q23.3	*B4GALT3*, *ADAMTS4*	123
1q24.2	*CD247*	123
1q25.1	*TNFSF18*, *TNFSF4*	123
1q31.3	*DENND1B*, *CRB1*	129
1q32.1	*ADORA1*	123
2p25.1	*ID2*	123
2q12	*IL18R1*, *IL1RL1*, *IL1RL2*	108, 123, 124, 137
2q37.3	*D2HGDH*	123
3q28	*LPP*	123
4p14	*TLR1*	123
4q31.21	*LOC729675*	63
5q12.1	*PDE4D*	62
5q22.1	*SLC25A46*, *TSLP*	63, 123, 137
5q31	*RAD50*, *NDFIP1*	123
6p21.31	*GRM4*, *HGMA1*	4
6p21.32	*PBX2*, *NOTCH4*, *C6orf10*, *BTNL2*, *HLA-DRA*, *HLA-DRB5*, *HLA-DQA1*, *HLA-DQ*, *HLA-DQB1*, *HLA-DQA2*, *HLA-DOA*, *HLA-DPB1*	63, 80, 108, 113, 123, 124
6p21.33	*HLA-C*, *MICA*	123
6q15	*BACH2*	123
7q22.3	*CDHR3*	123
8q21.13	*TPD52*, *ZBTB10*	123
8q24.11	*SLC30A8*	113
9p21.2	*EQTN*, *TEK*, *MOB3B*	4
9p24.1	*RANBP6*, *IL33*	108, 123, 137
9q34.11	*PTGES*	4
10p14	*LOC338591*, *GATA3*	63, 123
10q21.3	*JMJD1C*	4
11q13.5	*C11orf30*, *LRRC32*	45, 123
12q13.2	*CDK2*, *IKZF4*	63
12q13.3	*STAT6*, *LRP1*	123
14q24.1	*RAD51B*	123
15q22.2	*RORA*	123
15q22.33	*SMAD3*	108, 123
16p13.13	*CLEC16A*	123
17q12–21	*GRB7*, *IKZF3*, *ZPBP2*, *GSDMB*, *ORMDL3*, *GSDMA*	108, 109, 123, 137, 145, 156
22q12.3	*IL2RB*	108

This table is collated from the publicly available GWAS Catalog (19), listing all GWAS hits for asthma per se (excluding intermediate or compound phenotypes) with a p value $< 5 \times 10^{-8}$ at the time of writing.

Downloaded from www.annualreviews.org. Guest (guest) IP: 24.117.150.43 On: Tue, 14 Oct 2025 03:05:08

Other loci achieving significant *p* values (i.e., below the genome-wide significance threshold of 5×10^{-8}) in multiple GWASs include a site on chromosome 2q12 in the vicinity of several interleukin receptor genes, namely *IL1RL1*, *IL1RL2*, and *IL18R1*. While these multigene loci are challenging to dissect, it is notable that *IL1RL1* encodes the receptor for IL33. The gene that encodes IL33 is separately implicated in the genetic etiology of asthma through a replicated locus on chromosome 9p24.1. IL33, a transcription factor and cytokine, is expressed in the airway epithelium, where it acts as a proinflammatory danger signal following its release from cells undergoing necrosis (resulting from infection or damage). Other widely observed sites of association include a region on 5q22.1 that contains the mitochondrial solute carrier gene *SLC25A46* and the hemopoietic cytokine gene *TSLP*, a complex region located within the major histocompatibility locus (6p21.32), the *C11orf30-LRRC32* site implicated elsewhere in an array of atopic traits (11q13.2), and the signal transducer and transcriptional modulator gene *SMAD3* (15q22.33).

A notable observation from GWASs of asthma is the profound lack of overlap between the results of early low-resolution linkage and positional cloning studies and those of subsequent high-resolution GWASs. Examination of the GWAS Catalog (19) shows that no marker positioned in or around the eight positionally cloned genes (**Table 1**) has yet achieved genome-wide significance for asthma, or even a more modest threshold of $p \leq 5 \times 10^{-6}$. Markers proximal to these genes do, however, achieve moderately low *p* values for a spectrum of alternative traits, including several of possible relevance to asthma. These traits include indices related to obesity (*DPP10*), inflammatory diseases of alternative epithelial surfaces (*HLA-G*), measures of lung function (*CYFIP2* and *NPSR1*), atopy (*HLA-G*), infection, and host–organism interactions (*DPP10* and *HLA-G*) (**Table 1**).

There are various explanations for this lack of overlap. Standard GWASs are not powered to capture linkage disequilibrium between typed variants and untyped variants of low minor allele frequency, while classic linkage studies are capable of detecting rare variants but require a larger effect size (high penetrance) to permit detection. Individual linkage signals may also reflect allelic and/or locus heterogeneity, with contrasting alleles or even genes underlying the signal in different pedigrees. These differences mean that the two approaches may target partially separable components of the genetic architecture and as such may account for an apparent lack of communality between the results of linkage studies and those of GWASs in asthma. However, the failure of GWASs to detect signals in regions implicated by linkage analysis may also suggest that the linkage represents a false positive, particularly when imputation of rarer untyped variants is included in the GWAS.

Asthma and Related Phenotypes

Epidemiological studies have shown that asthma commonly co-occurs at the level of the individual, or more broadly at the level of the immediate family unit, with a diverse spectrum of traits, from gastroesophageal reflux disease to obesity. These patterns of co-occurrence suggest a level of overlap in the underlying mechanisms of pathogenesis.

The atopic triad. Perhaps the most widely recognized pattern of co-occurrence is the one of asthma, atopic dermatitis (eczema), and allergic rhinitis (hay fever), which together are referred to as the atopic triad and characteristically present clinically in a temporal sequence known as the atopic march. Within this sequence, atopic dermatitis is typically the first component to manifest, with approximately 20–30% of individuals with mild disease and 70% of those with severe disease going on to develop asthma. Individuals who undergo this distinctive sequence of disease progression frequently exhibit a more severe and persistent phenotype, with increased risk of allergen sensitization.

Downloaded from www.annualreviews.org. Guest (guest) IP: 24.117.150.43 On: Tue, 14 Oct 2025 03:05:08

Downloaded from www.annualreviews.org. Guest (guest) IP: 24.117.150.43 On: Tue, 14 Oct 2025 03:05:08

All three of the atopic triad traits are strongly heritable (asthma, 35–95%; allergic rhinitis, 33–91%; and atopic dermatitis, 71–84%) (116) and show patterns of familial aggregation. The traits demonstrate marked (though incomplete) genetic correlations, indicating the presence of both shared and unique genetic influences (asthma–atopic dermatitis, 0.35–0.55; asthma–allergic rhinitis, 0.47–0.90; allergic rhinitis–atopic dermatitis, 0.62–0.73) (86, 140). The molecular mechanisms for this multi-morbidity are unclear. Comparison of susceptibility loci revealed by early linkage studies and subsequent single-trait GWASs provided little evidence of overlap. Indeed, a greater level of coincidence was reported between atopic dermatitis and the Th1-driven skin disease psoriasis, although this was antagonistic in nature (mapping to shared genes but opposing alleles) (148). Similar instances of inverse (antagonistic) association have been documented across a variety of disease pairings, from chronic obstructive pulmonary disease and pulmonary fibrosis (65) to inflammatory bowel disease (IBD) and type 1 diabetes (146), suggesting the possible presence of balancing selection at these genetic sites. Pleiotropic effects, whereby a single locus influences two or more traits, have been widely reported for atopic dermatitis. A recent GWAS observed pervasive sharing of loci (121). Almost 78% of atopic dermatitis–associated loci identified were implicated in alternative immune-mediated traits, from IBD to psoriasis and multiple sclerosis. While none of these genome-wide significant loci matched established sites of association to asthma per se, the authors reported that a third of established asthma loci achieved nominal significance levels in atopic dermatitis with an equivalent direction of effect. This suggests the possibility of small effect sizes in asthma that require substantial levels of power for detection at a genome-wide significance level.

More recent attempts to deconstruct the molecular mechanisms of the atopic triad multimorbidity have adopted two main approaches. The first is in silico, involving data mining and the construction of protein-level functional interaction networks. This strategy has shown that atopic triad diseases share a larger fraction of proteins and a greater level of interconnectedness than would be expected by chance (2). Specifically, five proteins (IL4, IL13, IL1RL1, IL18R1, and TSLP) and two cellular pathways (involving GATA3-related and IL4 signaling) were identified as common to all three atopic triad traits. *GATA3* is a recognized regulator of T cell development, while *TSLP* encodes thymic stromal lymphopoietin, a hemopoietic cytokine that promotes Th2 immune responses. Together, these data highlight a common theme of the type 2 response.

The second approach has involved conducting cross-trait GWASs using compound atopic triad phenotypes. Marenholz et al. (97) examined infantile atopic dermatitis followed by childhood asthma and identified seven contributory loci: 1q21.3 (*FLG*), 5q31.1 (*IL5/KIF3A*), 6p12.3 (*EFHC1*), 11q13.1 (*AP5B1/OVOL1*), 11q13.5 (*C11orf30/LRRC32*), 12q21.3 (*SLC6A15/TMTC2*), and 17q21 (*IKZF3*). Subsequently, Ferreira et al. (46) extended this approach in a substantial collection of more than 360,000 individuals, seeking genetic risk factors shared among all three atopic triad traits. The authors identified 136 independent risk variants underlying this shared risk, more than a third of which mapped to novel loci. Only a small minority of variants (4.4%) showed stronger effects in one atopic trait relative to the remaining two. These included *GSDMB*, located in the major chromosome 17q12–21 asthma locus (with risk variants relatively more common in asthma), and *FLG*, the most significant and replicable site of genetic association with atopic dermatitis to date (more common in atopic dermatitis).

Filaggrin, encoded by *FLG*, forms a key structural component of the terminally differentiated epidermis. The gene contains a spectrum of loss-of-function mutations that associate with asthma occurring in conjunction with atopic dermatitis but not asthma occurring independently of atopic dermatitis (118). These data alongside an absence of *FLG* expression in the airway epithelia (excepting the hair-bearing outer nasal cavity) suggest that, in the subset of cases bearing *FLG* mutations, asthma may be a secondary to a compromised, "leaky" skin barrier that allows percutaneous transfer of allergens and pathogens. Consistent with this hypothesis, mice naturally deficient for filaggrin (the so-called flaky tail mutants) show increased levels of percutaneous allergen priming (43), and in humans, *FLG* mutations are additionally associated with increased rates of several Th2 phenotypes (such as food allergy) that manifest at sites anatomically remote to the epidermis (70).

Obesity. In addition to atopic traits, there is strong epidemiological evidence of an association between asthma and obesity, particularly in affluent countries. Obesity precedes the development of asthma, conferring a dose-dependent effect on asthma incidence (150). No consistent relationship, however, has been identified between excess weight and atopic status (125, 149), leading some to posit obese asthma as a separable disease phenotype (93). Proposed explanations for the epidemiological relationship between asthma and obesity include shared co-morbidities (such as gastroesophageal reflux disease) and/or shared pathological features (such as systemic inflammation, oxidative stress, or leptin resistance). The adipose-derived hormone leptin is a key metabolic signal and is typically observed at relatively increased levels in obesity. Leptin is associated with several immune functions and regulates airway diameter independently of inflammation (7), which is consistent with the relative inefficacy of anti-inflammatory agents in the treatment of obese asthma (122).

Asthma and obesity possess a modest genetic correlation of 0.29, accounting for approximately 8% of the total genetic component of obesity (59). The genetic underpinnings of this joint susceptibility are largely undefined, with a single exception: a common inversion on chromosome 16p11.2 (55) that affects the expression of a collection of genes participating in the regulation of energy balance and immunity. Analysis of GWAS signals relating to body mass index or asthma in asthmatics (a GWAS-informed candidate gene strategy) did not reveal any significant pattern of overlap (105). Likewise, attempts to map the shared genetic origins of asthma and obesity directly through GWASs have proven relatively unsuccessful. A GWAS of body mass index in 23,000 asthmatics yielded only limited evidence of a heterogeneous effect of *DENND1B* [an overall GWAS hit for asthma (129); see **Table 1**] in asthmatic children, an effect that could not be replicated outside of the discovery cohort (104). Equally, a GWAS designed to model the interaction between asthma and body mass index on a genome-wide scale identified no variants significant at a genome-wide threshold, although multiple variants in the classic asthma 17q21 locus were classified as suggestive (147).

Chronic obstructive pulmonary disease. Considering the adult age group, asthma can also overlap with chronic obstructive pulmonary disease (133). This disease is characterized by a progressive, refractory airflow limitation, typically associated with long-term tobacco smoke exposure. A genetic correlation of 0.38 is defined between chronic obstructive pulmonary disease and asthma (65). With the exception of the locus on chromosome 16p11.2, previously implicated in the joint susceptibility to asthma and obesity (55), GWASs have not yet revealed shared sites of association between asthma and chronic obstructive pulmonary disease (65).

Inflammatory bowel disease. Several parallels can be drawn between asthma and IBD. IBD encompasses two major clinical entities: Crohn's disease and ulcerative colitis. Both IBD and asthma represent chronic relapsing inflammatory phenotypes that affect mucosal epithelial surfaces, albeit remotely from one another (in the gastrointestinal and respiratory tracts, respectively). The diseases additionally share several risk factors. As with asthma, a lack of breastfeeding is associated with an increased risk of IBD, as is early exposure to antibiotics. Furthermore, there is now accumulating evidence of an elevated prevalence of asthma (and atopic dermatitis) in IBD, with a

Downloaded from www.annualreviews.org. Guest (guest) IP: 24.117.150.43 On: Tue, 14 Oct 2025 03:05:08

diagnosis of asthma associated with significantly increased odds of early- and late-onset ulcerative colitis as well as Crohn's disease (76). Familial clustering of IBD is consistent with a genetic component, with pooled twin studies indicating a heritability of 0.75 for Crohn's disease and 0.67 for ulcerative colitis (56). While the genetic correlation between asthma and IBD is not, to the best of our knowledge, currently defined, evidence supporting the presence of genetic overlap dates back to early linkage studies (26).

More recent attempts to map the genetic determinants of IBD (IBD overall or its two major clinical entities separately) through GWASs confirmed genetic intersection between these traits, identifying at least 11 sites of association ($p < 5 \times 10^{-8}$) coincident with those described by asthma GWASs: chromosomes 1q25.1 (*TNFSF18* and *TNFSF4*) (13, 49, 72), 1q31.3 (*DENND1B* and *CRB1*) (32, 49, 72, 87), 2q12 (*IL18R1*, *IL1RL1*, and *IL1RL2*) (32, 49, 72, 87), 3q28 (*LPP*) (32), 5q31 (*NDFIP1*) (32, 49, 72, 87), 6p21.32 (*PBX2*, *NOTCH4*, *C6orf10*, *BTNL2*, *HLA-DRA*, *HLA-DRB5*, *HLA-DQA1*, and *HLA-DQA2*) (5, 8, 32, 47–49, 72–74, 77, 87, 102, 128, 138, 155, 158, 159), 6p21.33 (*HLA-C* and *MICA*) (32, 72, 73), 6q15 (*BACH2*) (32, 49, 72, 87), 11q13.5 (*C11orf30*) (5, 13, 32, 49, 72, 87), 15q22.33 (*SMAD3*) (32, 49, 72, 87), and 17q12–21 (*GRB7*, *IKZF3*, *ZPBP2*, *GSDMB*, *ORMDL3*, and *GSDMA*) (5, 13, 32, 49, 72, 87, 102) (**Table 2**). Among these sites, the 17q12–21 locus has received considerable research attention because it is the most significantly associated locus for asthma; shows replicable association with ulcerative colitis, Crohn's disease, and IBD overall; and coincides with the most significant Crohn's disease–associated eQTL (13). Other loci of interest include *SMAD3*, which encodes SMAD family member 3, a signal and transcriptional modulator that is responsive to transforming growth factor β (TGF-β) and is therefore key to the TGF-β-mediated epithelial-to-mesenchymal transition and fibrosis. Similarly, *BACH2* (BTB domain and CNC homolog 2) encodes a transcription factor essential for T and B lymphocyte differentiation. *C11orf30* encodes the nuclear protein EMSY, which has postulated roles in epithelial immunity, growth, and/or differentiation (41). *C11orf30* has also been implicated in a range of atopy-related traits, from polysensitization to atopic dermatitis, eosinophilic esophagitis, and total serum IgE. Together, these data indicate substantial levels of genetic communality between IBD and asthma and direct emphasis toward the mucosal immune barrier in these diseases.

Asthma and Infection

A role of early severe viral respiratory infections in the development and exacerbation of asthma in children is widely recognized, particularly with regard to the respiratory pathogens rhinovirus and respiratory syncytial virus (52). Both of these viruses are early predictors of recurrent wheezing, with early rhinovirus-related wheeze predictive of subsequent asthma (88). The mechanism for this phenomenon, however, is poorly defined. Evidence of host genetic factors in respiratory tract infectious diseases is controversial, marred by issues of bias and confounding (120). A recent meta-analysis of 386 studies targeting infectious diseases of the respiratory tract nonetheless highlighted a role for the *IL4-TLR2-CCL2* axis, with *IL4* sequence variants achieving significance in a pooled model, indicating shared host effects common to multiple respiratory infections (120). Notably, sequence variants proximal to *IL4* have previously been implicated in the etiology of asthma through candidate gene association studies (see above). Mice deficient for *IL4* show increased susceptibility to some respiratory pathogens as well as increased mortality (110), consistent with the gene's role in host–pathogen dynamics. A direction of focus onto the mucosal barrier arising from GWASs adds weight to the rationale for investigating viral and other microbial infections in asthma and underscores the importance of ensuring that such data are captured efficiently at the time of study population collection.

Downloaded from www.annualreviews.org. Guest (guest) IP: 24.117.150.43 On: Tue, 14 Oct 2025 03:05:08

EPIGENOME-WIDE ASSOCIATION STUDIES

The term epigenetics describes changes in gene activity and/or function that do not involve alteration of the DNA sequence. DNA methylation was the first epigenetic change to be widely studied, following the observation that CpG methylation is associated with gene silencing and the patterning of gene expression that determines cellular types and functions (33). CpG sites represent positions in the DNA sequence in which a cytosine nucleotide and a guanine nucleotide are separated by a single phosphate group when read in the 5′-to-3′ direction (5′-C-p-G-3′). CpG islands encompass broader regions of DNA that contain a relative enrichment of CpG sites. Epigenetic investigations also include modifications of the histones and other regulatory elements that bind to DNA [exemplified in the Encyclopedia of DNA Elements (ENCODE) project].

While the original coinage of the term epigenetics implied epigenetic inheritance (e.g., at loci showing genomic imprinting), the vast majority of epigenetic modifications are not heritable. In general, epigenetic alterations can be considered responses to environmental influences, providing exposure-dependent plasticity in gene activity or function. Although epigenetically modified genes and pathways may be central to disease pathobiology, they are likely to represent changes in gene regulation in response to environmental or other factors and are not causal of disease susceptibility in the same way as true genetic variants.

Further factors that should be considered in interpreting epigenome-wide association studies (EWASs) are the distinctive epigenetic patterns of particular cell types and the cellular heterogeneity present in peripheral blood and different tissues (84). In contrast to genetic studies, where polymorphisms are a constant feature of DNA derived from any tissue of the human body, epigenetic modifications contain strong cell-specific signatures. Consequently, for example, peripheral-blood leukocytes may not contain core epigenetic events that are present in asthmatic airways or eczematous skin.

While it has been suggested that generalized abnormalities of methylation may be present in asthma (12), the hypothesis has subsequently not gained support. Well-powered EWASs that have controlled for cellular heterogeneity have, however, generated some consistent and interesting results that indicate the importance of particular pathways and cell types in the etiology of atopic asthma.

Epigenetic Features of Asthma: Outside of the Lung

In 2015, Liang et al. (85) surveyed epigenetic associations between serum IgE concentrations and methylation at loci concentrated in CpG islands in 95 nuclear pedigrees, using a first-generation Illumina methylation array to test DNA from peripheral-blood leukocytes. They validated positive results in additional families and a general population sample. Replicated associations (false discovery rate $<10^{-4}$) were found between IgE and low methylation at 36 loci. Many genes annotated to these loci encode known eosinophil products (in particular *IL5RA*, encoding the eotaxin receptor), whereas other associations implicated phospholipid inflammatory mediators (such as the one encoded by *LPCAT2*), specific transcription factors, and mitochondrial proteins (such as the one encoded by *L2HGDH*).

Modeling of the association with differential white cell counts showed that 10% of the variation in IgE level was attributable to eosinophil counts and that the top three loci (indicating the level of eosinophil activation) accounted for a further 13% of IgE variation. The authors confirmed eosinophil activation by showing that methylation in isolated eosinophils differed significantly in subjects with and without asthma and high IgE levels.

Downloaded from www.annualreviews.org. Guest (guest) IP: 24.117.150.43 On: Tue, 14 Oct 2025 03:05:08

A subsequent study of approximately 1,000 children from the Avon Longitudinal Study of Parents and Children assessed associations of asthma and wheeze with DNA methylation at 450,000 sites (using an Illumina methylation array), extending the tests for differentially methylated regions beyond CpG islands. Samples were taken at 7.5 and 16.5 years of age (6). The authors identified 302 CpGs associated with current asthma status (false discovery rate–adjusted p value $<$ 0.05) and 445 associated with current wheeze status at 7.5 years, with substantial overlap between the two. Genes annotated to the 302 associated CpGs were related to IL4 production and eosinophil migration. As with the adult EWAS (85), the associations were attenuated when adjusted for eosinophil and neutrophil cell count estimates. At 16.5 years, two sites, mapping to the *AP2A2* and *IL5RA* genes, remained associated with current asthma after adjustment for cell counts.

Epigenome-wide DNA methylation was most recently investigated in the whole-blood DNA of asthma cases and controls at ages 4–8 years from four European birth cohorts within the Mechanisms of the Development of Allergy project (153). Cell type specificity was addressed in eosinophils and nasal epithelial cells. Consistently lower methylation levels were observed at 14 loci that showed genome-wide significance in meta-analyses; all of the loci were associated across childhood from ages 4 to 16 years but not at birth. All 14 CpGs were significantly associated with asthma in a replication study using whole-blood DNA and were strongly associated with asthma in purified eosinophils. Whole-blood transcriptional signatures associated with these CpGs indicated increased activation of eosinophils, effector/memory CD8-T, and natural killer cells and reduced naive T cells. Of the 14 CpGs, 5 were associated with asthma in respiratory epithelial cells, indicating cross-tissue epigenetic effects. These CpGs and their associated transcriptional profiles suggest activation of eosinophils and cytotoxic T cells in childhood asthma, consistent with early-life microbial exposures.

A smaller study of methylation in cord blood (36) found that methylation in *SMAD3* was selectively increased in asthmatic children of asthmatic mothers and was associated with childhood asthma risk. As highlighted above, GWASs of asthma have found *SMAD3* to be important, and further analyses may be necessary to separate the epigenetic changes from the effects of heritable polymorphism.

EWAS association between genome-wide DNA methylation in whole-blood cells and total IgE levels has been examined in two panels of Hispanic children (22). The results confirmed many associations with total IgE levels previously reported in Caucasians. As with other investigations, adjustment for whole-blood cell types resulted in markedly fewer significant sites. Top novel findings from an adjusted meta-analysis were in the genes *ZFPM1* ($p = 1.5 \times 10^{-12}$), *ACOT7* ($p = 2.5 \times 10^{-11}$), and *MND1* ($p = 1.4 \times 10^{-9}$).

Overall, the above EWASs have identified consistent sets of markers that show activated eosinophils to be of particular importance in childhood asthma and adult asthma when accompanied by a high serum IgE level. Eosinophils develop in the bone marrow and are activated and recruited into tissues in response to stimuli that include IL5 and eotaxin (126). IL5 and its specific receptor subunit, IL5RA, are therapeutic targets for asthma, and these epigenetic findings validate drugs that include benralizumab (which binds to IL5RA) and the humanized IL5-specific antibodies mepolizumab and reslizumab. These biologics are very expensive (on the order of US$30,000 per patient each year), and as such, assays for measuring the level of activation of IL5RA may become of value in predicting which patients will respond to treatment. Alternatively, the *LPCAT2* gene, implicated by Liang et al. (85), may define a way to treat asthma with small-molecule antagonists that will prevent the stimulus-dependent formation of the potent proinflammatory lipid mediator platelet-activating factor (PAF) (127).

Downloaded from www.annualreviews.org. Guest (guest) IP: 24.117.150.43 On: Tue, 14 Oct 2025 03:05:08

Epigenetic Features of Asthma: Within the Lung

A study of the epigenome in primary airway epithelial cells from 74 asthmatic and 41 non-asthmatic adults (111) revealed a regulatory locus on chromosome 17q12–21 (the same locus identified by asthma GWASs) associated with asthma risk and epigenetic signatures of specific asthma endotypes. *ORMDL3* expression was related to the differentially methylated region at this locus, while the expression of *GSDMB* was not. A differentially methylated region near *CCL26* (encoding eotaxin 3) also had a significantly associated eQTL. This important study reminds us that asthma is primarily a disease of the airways, whereas effects in peripheral-blood leukocytes reflect changes in the circulating immune system. This may explain why the peripheral-blood leukocyte studies outlined above primarily reported strong associations with the total serum IgE, with relatively weak findings for asthma itself. Despite the technical challenges of obtaining samples from within the lung, further genomic and methylation studies of the airway epithelium appear necessary to enable the field to move forward.

TRANSCRIPTOMICS

The development of array-based and, subsequently, RNA sequencing technologies has fostered high-throughput transcriptomic analyses of asthma across a host of cell populations, time frames, and exposure conditions.

Transcriptomics as a Tool for Defining the Impact of Disease on a Tissue or Cell System

In one of the first large-scale, genome-wide gene expression studies of asthma, Dixon et al. (38) characterized the transcriptomes of peripheral-blood leukocyte–derived lymphoblastoid cell lines of children with and without asthma and examined the relationship between transcript abundance and genotype. Clear patterns of heritability were observed, with more than 15,000 transcripts demonstrating a narrow-sense (h_2) heritability greater than 0.3, including both local and distal effects and clear patterns of enrichment for specific biological processes. Nonetheless, only 10 transcripts showed significant disease-related differences in abundance, and none of these retained significance following adjustment for multiple comparisons. This result is perhaps unsurprising given the context in which disease-related differences in gene expression were sought—i.e., unchallenged, immortalized cultured cell lines. Nevertheless, the authors were able to confirm that asthma-associated markers in the 17q23 region account for 29.5% of variance in *ORMDL3* expression, supporting *ORMDL3* as a major candidate gene in this region.

Transcriptional signatures of asthma have since been sought in a variety of minimally invasive disease-relevant tissues, ranging from sputum (157) to whole blood (16, 27) and bronchial brushings (91). Peripheral-blood data sets suggest the presence of replicable transcriptomic signatures relating to the extent of asthma control, with the strongest and most highly reproducible enriched gene sets expressed by specific immune cell lineages, including 11 eosinophil-specific gene sets (suggesting parallels with EWAS data) (27). Whole-blood results also indicate variation in gene expression by disease severity, with a substantially greater magnitude of differential expression documented between healthy subjects and severe asthmatics than between healthy subjects and those with mild or moderate disease. However, similar gene sets have been reported as affected in both comparisons, suggesting a disease-severity-related transcriptomic continuum (16). That said, in their 2016 analysis of the Unbiased Biomarkers in Prediction of Respiratory Disease Outcomes (U-BIOPRED) cohorts, Bigler et al. (16) showed that the number of differentially expressed genes

Downloaded from www.annualreviews.org. Guest (guest) IP: 24.117.150.43 On: Tue, 14 Oct 2025 03:05:08

in whole blood was dramatically cut (by as much as 90%) when blood cell counts were included as covariates, suggesting that the majority of differential expression may be accounted for by disease-related changes in peripheral-blood composition.

Shifting the compartment of study from peripheral blood to the epithelial barrier, in 2017 Loffredo et al. (91) performed a comparative analysis of publicly available transcriptomic data sets of asthma from bronchial epithelial cells. The authors identified a consensus signature of 159 core genes differentially expressed by disease status with a consistency in directionality across the majority of studies (five of the eight studies analyzed). Functional pathway analysis of these genes revealed a close-knit interaction network, with key features highlighting themes of epithelial–mesenchymal imbalance, remodeling, and dedifferentiation. Traditional markers of epithelial-to-mesenchymal transition were notably absent; instead, *EFNB2*, *FGFR1*, *FGFR2*, *INSR*, *IRS2*, *NOTCH2* (participating in ephrin, insulin, and Notch signaling), *NTRK2*, and *TLE1* were identified as central components of the interaction network. Key genes were differentially expressed across the full spectrum of disease severity and showed progression along the disease-severity gradient. Moreover, similar gene expression patterns could be approximated in organotypic air–liquid interface culture conditions via insulin depletion. These gene expression data support a hypothesis of barrier dysfunction in asthma and suggest the possibility that this dysfunction may be secondary to systemic factors (such as insulin insufficiency or resistance).

Transcriptomics as a Tool for Pathway or Candidate Gene Exploration

We have highlighted above that transcriptomics offers an unbiased method of characterizing pathways and their composite transcripts perturbed in the context of asthma, as well as cross-tissue and inter-subgroup comparisons. In addition, however, transcriptomics offers a valuable opportunity to catalog the impact of experimentally manipulated endogenous or exogenous factors on a cellular system, or to simply track gene expression patterns through key developmental stages or processes. Widely applied experimental manipulations include controlled knockdown of a single candidate gene, exposure to a known risk factor (e.g., microbial stimuli or organic or nonorganic airborne particulates), or a combination of the two.

By way of example, the process of class-switch recombination from IgA/IgG/IgM to IgE is of central importance to atopic diseases such as asthma, since it is required for B cells to express IgE. In 2016, Zhang et al. (161) monitored global gene expression and exon retention during the activation of immunoglobulin class switching in human B cells. The authors identified a complex yet organized transcriptional cascade accompanying the process of class-switch recombination, including genes known to participate in the moderation of circadian rhythms (*BHLHE40* and *NFIL3*). The involvement of clock genes is notable since circadian oscillations represent a prominent feature of asthma symptoms (39), and experimental deletion of an alternative clock gene (*BMAL1*) recently revealed links between the circadian clock and both viral airway pathology and asthma phenotypes (40).

The role of viral respiratory infection in asthma is a major area of emerging interest (see the section titled Asthma and Infection). To this end, Etemadi et al. (42) recently profiled the transcriptional response to rhinovirus infection in human alveolar epithelial cells. These data, derived from a controlled in vitro system, revealed a substantial impact of rhinovirus infection on global gene transcription, with distinct functional categories showing enrichment at the various stages of the infection time course, including an early inflammatory chemokine response. The most strikingly upregulated gene across all time points was the transcriptional regulator gene *EGR1* (early growth response 1), implicated elsewhere in the transcriptional response to smoke exposure in the nasal mucosa (99).

Downloaded from www.annualreviews.org. Guest (guest) IP: 24.117.150.43 On: Tue, 14 Oct 2025 03:05:08

Considering the process of lung development, Melén et al. (106) queried publicly available gene expression data sets that had assayed gene expression in the developing human or murine lung, and sought overlap with known asthma candidate genes. Candidate genes were not significantly overrepresented among developmentally regulated genes, but a small minority of candidates did demonstrate differential expression and consistent temporal trends in both mice and humans. These included *NOD1*, *EDN1*, and *CCL5*, as well as the positionally cloned *HLA-G* and *RORA* genes, both of which have more recently been identified to be of importance by asthma GWASs. These data cumulatively add context to existing associations and suggest possible roles in lung development or morphogenesis.

THE LUNG MICROBIOME

It has long been hypothesized that microbiotas play a key role in the etiology of asthma and, more broadly, atopic disease. The prevalence of asthma has increased dramatically over recent decades in parallel with a decline in the rates of infectious disease. These changes accompany improvements in sanitation, cleanliness, and vaccination. Small family sizes, birth by caesarian section, and urban living all represent risk factors for asthma, while early farm exposures and breastfeeding confer protective effects. Such observations have been assimilated into the hygiene hypothesis, first set out in 1989 (136), positing that reduced early microbial exposure and its impacts on immunity underlie the post–Industrial Revolution atopy and asthma epidemic. Responsible for a transformation in our understanding of microbial factors in asthma has been a revolution of a different kind. Only approximately 5% of bacteria can be cultured in the laboratory through standard techniques—the so-called great plate count anomaly (135). By contrast, next-generation sequencing approaches, which make use of the balance between sequence conservation and variation in the bacterial 16S rRNA gene, enable the isolation and accurate identification of most microorganisms. The technique has been successfully applied to the upper and lower airways as well as the gut and has provided novel insights into the role(s) played by bacteria in asthma.

Historically, the lung was considered a sterile landscape. Application of culture-independent techniques has, however, shown that neither the upper nor the lower airways can be considered sterile in either health or disease (asthma, chronic obstructive pulmonary disease, or idiopathic pulmonary fibrosis) (reviewed in 95). Moreover, accumulating evidence suggests that colonization of some anatomical sites may even precede birth, being initiated in utero (1, 23). In health, microbiotas detected in the lung demonstrate considerable overlap with those seen in the upper airways (14, 98), arguably consistent with microaspiration as a source of lung microbiotas. In disease, the overlap between the upper airways and lung is less clear (37, 61).

Culture-independent microbiome profiling specifically in asthma has revealed significant differences between cases and controls in the lower airways, in particular with regard to members of the gram-negative phylum Proteobacteria (54, 61). Interestingly, *Haemophilus parainfluenzae*, also a member of the Proteobacteria phylum, appears to promote blunt corticosteroid responses through activation of the TAK1/MAPK pathways, suggesting a possible microbial mechanism for corticosteroid resistance in asthma (54). While a role for Proteobacteria in asthma is contentious, data from murine models also highlight weak inflammatory properties of airway commensals (accounting for their tolerance), while asthma-associated Proteobacteria trigger lung inflammation and associated pathology (79).

CONCLUSIONS: A GENOMIC VIEW OF ASTHMA

The field of asthma genetics has changed dramatically over the last decade. Technological advancements have brought with them the tools with which to interrogate the human genome,

Downloaded from www.annualreviews.org. Guest (guest) IP: 24.117.150.43 On: Tue, 14 Oct 2025 03:05:08

epigenome, and transcriptome, as well as the microbial flora of humans, with unprecedented accuracy and volume. This toolbox has facilitated the stepwise accrual of information regarding the pathogenic mechanisms of asthma, free from hypothesis or bias, and enabled direct comparison between diseases. The challenge now is twofold: first, generating a meaningful synthesis of these data, which are both substantial and diverse, and second, explaining the considerable proportion of asthma heritability that remains unaccounted for despite the volume of existing data—challenges that are closely interwoven.

At the time of writing, GWASs alone have yielded at least 38 genomic loci that show association with asthma with p values that meet or exceed the widely applied genome-wide threshold for significance of 5×10^{-8}. These loci show little to no correspondence with the eight previously positionally cloned asthma susceptibility genes, and likewise, historical candidate genes do not feature prominently in asthma GWAS results. The explanation for this lack of intersection may in part be statistical, reflecting the presence of false positives among reported susceptibility genes. This is not, however, the only possible explanation; also likely to be influential are a disparity in the types of effects that are amenable to capture by the techniques of linkage and association (including the potential role of rare variants) and the overarching role of unmeasured variation. The latter encompasses the phenotype itself, the underlying genetic architecture, and salient environmental exposures.

Epidemiological data have clearly demonstrated the relevance of a broad spectrum of environmental factors in the etiology of asthma, from tobacco smoke exposure to early respiratory infection and birth by caesarian section. More recent analyses of the lung microbiome in asthma have built on these observations, indicating a dysbiosis in disease and controversially highlighting a significant role for Proteobacteria. The participation of environmental factors in the etiology of asthma renders gene × environment interaction a plausible explanation for inconsistent replication between cohorts. Indeed, analyses of the major asthma locus on chromosome 17q12–21 have revealed evidence of interplay with several exogenous factors, including early respiratory infection (130) and early passive environmental tobacco smoke (17), the latter of which has been independently replicated (141). Since narrow-sense heritability estimates may be inflated by the presence of nonadditive effects (96), gene × environment interactions may plausibly contribute to the issue of apparent missing heritability. Today, gene × environment interactions are being modeled on a genome-wide scale through GWASs (143), but the success of such approaches will be tied to the selection and precise measurement of pertinent environmental exposures, the collection of information regarding their temporal sequence, and the generation of sufficient statistical power. Clarification of the relevant environmental influences in asthma may therefore be warranted, enabling more straightforward identification and replication of environmentally malleable genetic effects.

A second observation from the genetic studies of asthma has been the genetic correlation between asthma and a range of other phenotypes, and thence the accompanying overlap in susceptibility genes or variants. Analyses of this overlap have already afforded some insights into shared mechanisms, e.g., pointing to the mucosal immune barrier in IBD and the type 2 response in atopic triad diseases. Indeed, in some cases, such as the epidermal barrier gene *FLG*, association with asthma appears to occur only within the context of a second trait (in the case of *FLG* atopic dermatitis). Such observations underscore the importance of patient stratification and clear phenotype definition in study design and lend weight to the notion that asthma is both genetically and clinically heterogeneous. Unfortunately, in the majority of cases, the effect size conferred by any given risk variant is relatively low, limiting their individual utility for distinguishing patients in a clinical setting. Nevertheless, these variants and their patterns of pleiotropy provide valuable information on the mechanisms of disease causation and co-morbidity. Unlike the cancer field

Downloaded from www.annualreviews.org. Guest (guest) IP: 24.117.150.43 On: Tue, 14 Oct 2025 03:05:08

(notably breast cancer), the field of asthma remains at some distance from personalized medicine, but it appears likely that at least a subset of genes that affect asthma susceptibility also influence risk for other disease traits. Furthermore, transcriptomic data suggest that blood-based transcriptional signatures exhibit relationships with features of asthma itself, such as disease severity or control.

Transcriptomic screening provides a systematic technique for assessing the impact of a single gene (through experimental up- or downregulation) on a cellular system, and this will be key to understanding the roles of asthma susceptibility genes defined to date. Such techniques are already widely applied for cataloging the transcriptional underpinnings of disease-relevant processes, be it class-switch recombination in B cells or response to infection in alveolar epithelial cells. Moving toward a more complete understanding of the genetic basis of asthma will require characterizing the role(s) of established susceptibility genes through these and other similar techniques, as well as focusing on patient stratification and characterization of environmental influences.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

LITERATURE CITED

1. Aagaard K, Ma J, Antony KM, Ganu R, Petrosino J, Versalovic J. 2014. The placenta harbors a unique microbiome. *Sci. Transl. Med.* 6:237ra65
2. Aguilar D, Pinart M, Koppelman GH, Saeys Y, Nawijn MC, et al. 2017. Computational analysis of multimorbidity between asthma, eczema and rhinitis. *PLOS ONE* 12:e0179125
3. Allen M, Heinzmann A, Noguchi E, Abecasis G, Broxholme J, et al. 2003. Positional cloning of a novel gene influencing asthma from chromosome 2q14. *Nat. Genet.* 35:258–63
4. Almoguera B, Vazquez L, Mentch F, Connolly J, Pacheco JA, et al. 2017. Identification of four novel loci in asthma in European American and African American populations. *Am. J. Respir. Crit. Care Med.* 195:456–63
5. Anderson CA, Boucher G, Lees CW, Franke A, D'Amato M, et al. 2011. Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. *Nat. Genet.* 43:246–52
6. Arathimos R, Suderman M, Sharp GC, Burrows K, Granell R, et al. 2017. Epigenome-wide association study of asthma and wheeze in childhood and adolescence. *Clin. Epigenet.* 9:112
7. Arteaga-Solis E, Zee T, Emala CW, Vinson C, Wess J, Karsenty G. 2013. Inhibition of leptin regulation of parasympathetic signaling as a cause of extreme body weight-associated asthma. *Cell Metab.* 17:35–48
8. Asano K, Matsushita T, Umeno J, Hosono N, Takahashi A, et al. 2009. A genome-wide association study identifies three new susceptibility loci for ulcerative colitis in the Japanese population. *Nat. Genet.* 41:1325–29
9. Astle WJ, Elding H, Jiang T, Allen D, Ruklisa D, et al. 2016. The allelic landscape of human blood cell trait variation and links to common complex disease. *Cell* 167:1415–29
10. Autism Spectr. Disord. Work. Group Psychiatr. Genom. Consort. 2017. Meta-analysis of GWAS of over 16,000 individuals with autism spectrum disorder highlights a novel locus at 10q24.32 and a significant overlap with schizophrenia. *Mol. Autism* 8:21
11. Balaci L, Spada MC, Olla N, Sole G, Loddo L, et al. 2007. *IRAK-M* is involved in the pathogenesis of early-onset persistent asthma. *Am. J. Hum. Genet.* 80:1103–14
12. Barnes PJ. 2009. Targeting the epigenome in the treatment of asthma and chronic obstructive pulmonary disease. *Proc. Am. Thorac. Soc.* 6:693–96
13. Barrett JC, Hansoul S, Nicolae DL, Cho JH, Duerr RH, et al. 2008. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat. Genet.* 40:955–62

Downloaded from www.annualreviews.org. Guest (guest) IP: 24.117.150.43 On: Tue, 14 Oct 2025 03:05:08

14. Bassis CM, Erb-Downward JR, Dickson RP, Freeman CM, Schmidt TM, et al. 2015. Analysis of the upper respiratory tract microbiotas as the source of the lung and gastric microbiotas in healthy individuals. *mBio* 6:e00037
15. Bhatnagar P, Barron-Casella E, Bean CJ, Milton JN, Baldwin CT, et al. 2013. Genome-wide meta-analysis of systolic blood pressure in children with sickle cell disease. *PLOS ONE* 8:e74193
16. Bigler J, Boedigheimer M, Schofield JPR, Skipp PJ, Corfield J, et al. 2017. A severe asthma disease signature from gene expression profiling of peripheral blood from U-BIOPRED cohorts. *Am. J. Respir. Crit. Care Med.* 195:1311–20
17. Bouzigon E, Corda E, Aschard H, Dizier MH, Boland A, et al. 2008. Effect of 17q21 variants and smoking exposure in early-onset asthma. *N. Engl. J. Med.* 359:1985–94
18. Bouzigon E, Forabosco P, Koppelman GH, Cookson WOC, Dizier MH, et al. 2010. Meta-analysis of 20 genome-wide linkage studies evidenced new regions linked to asthma and atopy. *Eur. J. Hum. Genet.* 18:700–6
19. Burdett T, Hall PN, Hastings E, Hindorff LA, Junkins HA, et al. 2017. *GWAS Catalog: the NHGRI-EBI Catalog of published genome-wide association studies*. Accessed Nov. 22, 2017. **http://www.ebi.ac.uk/gwas**
20. Cancer Genome Atlas Netw. 2012. Comprehensive molecular portraits of human breast tumours. *Nature* 490:61–70
21. Chen J, Miller M, Unno H, Rosenthal P, Sanderson MJ, Broide DH. 2018. Orosomucoid-like 3 (ORMDL3) upregulates airway smooth muscle proliferation, contraction, and Ca^{2+} oscillations in asthma. *J. Allergy Clin. Immunol.* 142:207–18
22. Chen W, Wang T, Pino-Yanes M, Forno E, Liang L, et al. 2017. An epigenome-wide association study of total serum IgE in Hispanic children. *J. Allergy Clin. Immunol.* 140:571–77
23. Collado MC, Rautava S, Aakko J, Isolauri E, Salminen S. 2016. Human gut colonisation may be initiated in utero by distinct microbial communities in the placenta and amniotic fluid. *Sci. Rep.* 6:23129
24. Colodro-Conde L, Zhu G, Power RA, Henders A, Heath AC, et al. 2015. A twin study of breastfeeding with a preliminary genome-wide association scan. *Twin Res. Hum. Genet.* 18:61–72
25. Comuzzie AG, Cole SA, Laston SL, Voruganti VS, Haack K, et al. 2012. Novel genetic loci identified for the pathophysiology of childhood obesity in the Hispanic population. *PLOS ONE* 7:e51954
26. Cookson WOC, Moffatt MF. 2000. Genetics of asthma and allergic disease. *Hum. Mol. Genet.* 9:2359–64
27. Croteau-Chonka DC, Qiu W, Martinez FD, Strunk RC, Lemanske RF Jr., et al. 2017. Gene expression profiling in blood provides reproducible molecular insights into asthma control. *Am. J. Respir. Crit. Care Med.* 195:179–88
28. Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, et al. 2012. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* 486:346–52
29. Das S, Miller M, Beppu AK, Mueller J, McGeough MD, et al. 2016. GSDMB induces an asthma phenotype characterized by increased airway responsiveness and remodeling without lung inflammation. *PNAS* 113:13132–37
30. Davenport ER, Cusanovich DA, Michelini K, Barreiro LB, Ober C, Gilad Y. 2015. Genome-wide association studies of the human gut microbiota. *PLOS ONE* 10:e0140301
31. Davies ER, Kelly JF, Howarth PH, Wilson DI, Holgate ST, et al. 2016. Soluble ADAM33 initiates airway remodeling to promote susceptibility for allergic asthma in early life. *JCI Insight* 1:e87632
32. de Lange KM, Moutsianas L, Lee JC, Lamb CA, Luo Y, et al. 2017. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nat. Genet.* 49:256–61
33. Deaton AM, Webb S, Kerr AR, Illingworth RS, Guy J, et al. 2011. Cell type-specific DNA methylation at intragenic CpG islands in the immune system. *Genome Res.* 21:1074–86
34. Deliu M, Sperrin M, Belgrave D, Custovic A. 2016. Identification of asthma subtypes using clustering methodologies. *Pulm. Ther.* 2:19–41
35. Denham S, Koppelman GH, Blakey J, Wjst M, Ferreira MA, et al. 2008. Meta-analysis of genome-wide linkage studies of asthma and related traits. *Respir. Res.* 9:38
36. DeVries A, Wlasiuk G, Miller SJ, Bosco A, Stern DA, et al. 2017. Epigenome-wide analysis links *SMAD3* methylation at birth to asthma in children of asthmatic mothers. *J. Allergy Clin. Immunol.* 140:534–42

Downloaded from www.annualreviews.org. Guest (guest) IP: 24.117.150.43 On: Tue, 14 Oct 2025 03:05:08

37. Dickson RP, Erb-Downward JR, Freeman CM, McCloskey L, Beck JM, et al. 2015. Spatial variation in the healthy human lung microbiome and the adapted island model of lung biogeography. *Ann. Am. Thorac. Soc.* 12:821–30
38. Dixon AL, Liang L, Moffatt MF, Chen W, Heath S, et al. 2007. A genome-wide association study of global gene expression. *Nat. Genet.* 39:1202–7
39. Durrington HJ, Farrow SN, Loudon AS, Ray DW. 2014. The circadian clock and asthma. *Thorax* 69:90–92
40. Ehlers A, Xie W, Agapov E, Brown S, Steinberg D, et al. 2018. BMAL1 links the circadian clock to viral airway pathology and asthma phenotypes. *Mucosal Immunol.* 11:97–111
41. Esparza-Gordillo J, Weidinger S, Folster-Holst R, Bauerfeind A, Ruschendorf F, et al. 2009. A common variant on chromosome 11q13 is associated with atopic dermatitis. *Nat. Genet.* 41:596–601
42. Etemadi MR, Ling K-H, Zainal Abidin S, Chee H-Y, Sekawi Z. 2017. Gene expression patterns induced at different stages of rhinovirus infection in human alveolar epithelial cells. *PLOS ONE* 12:e0176947
43. Fallon PG, Sasaki T, Sandilands A, Campbell LE, Saunders SP, et al. 2009. A homozygous frameshift mutation in the mouse *Flg* gene facilitates enhanced percutaneous allergen priming. *Nat. Genet.* 41:602–8
44. Ferkol T, Schraufnagel D. 2014. The global burden of respiratory disease. *Ann. Am. Thorac. Soc.* 11:404–6
45. Ferreira MA, Matheson MC, Duffy DL, Marks GB, Hui J, et al. 2011. Identification of IL6R and chromosome 11q13.5 as risk loci for asthma. *Lancet* 378:1006–14
46. Ferreira MA, Vonk JM, Baurecht H, Marenholz I, Tian C, et al. 2017. Shared genetic origin of asthma, hay fever and eczema elucidates allergic disease biology. *Nat. Genet.* 49:1752–57
47. Franke A, Balschun T, Karlsen TH, Sventoraityte J, Nikolaus S, et al. 2008. Sequence variants in *IL10*, *ARPC2* and multiple other loci contribute to ulcerative colitis susceptibility. *Nat. Genet.* 40:1319–23
48. Franke A, Balschun T, Sina C, Ellinghaus D, Hasler R, et al. 2010. Genome-wide association study for ulcerative colitis identifies risk loci at 7q22 and 22q13 (*IL17REL*). *Nat. Genet.* 42:292–4
49. Franke A, McGovern DP, Barrett JC, Wang K, Radford-Smith GL, et al. 2010. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat. Genet.* 42:1118–25
50. Galfalvy H, Haghighi F, Hodgkinson C, Goldman D, Oquendo MA, et al. 2015. A genome-wide association study of suicidal behavior. *Am. J. Med. Genet. B* 168:557–63
51. Garcia-Etxebarria K, Bracho MA, Galan JC, Pumarola T, Castilla J, et al. 2015. No major host genetic risk factor contributed to A(H1N1)2009 influenza severity. *PLOS ONE* 10:e0135983
52. Garcia-Garcia ML, Calvo C, Ruiz S, Pozo F, Del Pozo V, et al. 2017. Role of viral coinfections in asthma development. *PLOS ONE* 12:e0189083
53. Glob. Asthma Netw. 2014. *The global asthma report 2014*. Rep., Glob. Asthma Netw., Auckland, N.Z.
54. Goleva E, Jackson LP, Harris JK, Robertson CE, Sutherland ER, et al. 2013. The effects of airway microbiome on corticosteroid responsiveness in asthma. *Am. J. Respir. Crit. Care Med.* 188:1193–201
55. Gonzalez JR, Caceres A, Esko T, Cusco I, Puig M, et al. 2014. A common 16p11.2 inversion underlies the joint susceptibility to asthma and obesity. *Am. J. Hum. Genet.* 94:361–72
56. Gordon H, Trier Moller F, Andersen V, Harbord M. 2015. Heritability in inflammatory bowel disease: from the first twin study to genome-wide association studies. *Inflamm. Bowel Dis.* 21:1428–34
57. Granada M, Wilk JB, Tuzova M, Strachan DP, Weidinger S, et al. 2012. A genome-wide association study of plasma total IgE concentrations in the Framingham Heart Study. *J. Allergy Clin. Immunol.* 129:840–45
58. Halapi E, Gudbjartsson DF, Jonsdottir GM, Bjornsdottir US, Thorleifsson G, et al. 2010. A sequence variant on 17q21 is associated with age at onset and severity of asthma. *Eur. J. Hum. Genet.* 18:902–8
59. Hallstrand TS, Fischer ME, Wurfel MM, Afari N, Buchwald D, Goldberg J. 2005. Genetic pleiotropy between asthma and obesity in a community-based sample of twins. *J. Allergy Clin. Immunol.* 116:1235–41
60. Hao K, Bosse Y, Nickle DC, Pare PD, Postma DS, et al. 2012. Lung eQTLs to help reveal the molecular underpinnings of asthma. *PLOS Genet.* 8:e1003029
61. Hilty M, Burke C, Pedro H, Cardenas P, Bush A, et al. 2010. Disordered microbial communities in asthmatic airways. *PLOS ONE* 5:e8578
62. Himes BE, Hunninghake GM, Baurley JW, Rafaels NM, Sleiman P, et al. 2009. Genome-wide association analysis identifies *PDE4D* as an asthma-susceptibility gene. *Am. J. Hum. Genet.* 84:581–93

Downloaded from www.annualreviews.org. Guest (guest) IP: 24.117.150.43 On: Tue, 14 Oct 2025 03:05:08

63. Hirota T, Takahashi A, Kubo M, Tsunoda T, Tomita K, et al. 2011. Genome-wide association study identifies three new susceptibility loci for adult asthma in the Japanese population. *Nat. Genet.* 43:893–96
64. Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K. 2002. A comprehensive review of genetic association studies. *Genet. Med.* 4:45–61
65. Hobbs BD, de Jong K, Lamontagne M, Bosse Y, Shrine NR, et al. 2017. Genetic loci associated with chronic obstructive pulmonary disease overlap with loci for lung function and pulmonary fibrosis. *Nat. Genet.* 49:426–32
66. Holt RJ, Vandiedonck C, Willis-Owen SAG, Knight JC, Cookson WOC, et al. 2015. A functional AT/G polymorphism in the 5′-untranslated region of *SETDB2* in the IgE locus on human chromosome 13q14. *Genes Immun*. 16:488–94
67. Howard R, Rattray M, Prosperi M, Custovic A. 2015. Distinguishing asthma phenotypes using machine learning approaches. *Curr. Allergy Asthma Rep.* 15:38
68. Ikari J, Inamine A, Yamamoto T, Watanabe-Takano H, Yoshida N, et al. 2014. Plant homeodomain finger protein 11 promotes class switch recombination to IgE in murine activated B cells. *Allergy* 69:223–30
69. Ioannidis JP, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis DG. 2001. Replication validity of genetic association studies. *Nat. Genet.* 29:306–9
70. Irvine AD, McLean WH, Leung DY. 2011. Filaggrin mutations associated with skin and allergic diseases. *N. Engl. J. Med.* 365:1315–27
71. Jiang Y, Zhang H. 2011. Propensity score-based nonparametric test revealing genetic variants underlying bipolar disorder. *Genet. Epidemiol.* 35:125–32
72. Jostins L, Ripke S, Weersma RK, Duerr RH, McGovern DP, et al. 2012. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* 491:119–24
73. Jung ES, Cheon JH, Lee JH, Park SJ, Jang HW, et al. 2016. *HLA-C*01* is a risk factor for Crohn's disease. *Inflamm. Bowel Dis.* 22:796–806
74. Juyal G, Negi S, Sood A, Gupta A, Prasad P, et al. 2015. Genome-wide association scan in north Indians reveals three novel HLA-independent risk loci for ulcerative colitis. *Gut* 64:571–79
75. Kenny EE, Pe'er I, Karban A, Ozelius L, Mitchell AA, et al. 2012. A genome-wide scan of Ashkenazi Jewish Crohn's disease suggests novel susceptibility loci. *PLOS Genet*. 8:e1002559
76. Kuenzig ME, Barnabe C, Seow CH, Eksteen B, Negron ME, et al. 2017. Athma is associated with subsequent development of inflammatory bowel disease: a population-based case-control study. *Clin. Gastroenterol. Hepatol.* 15:1405–12
77. Kugathasan S, Baldassano RN, Bradfield JP, Sleiman PM, Imielinski M, et al. 2008. Loci on 20q13 and 21q22 are associated with pediatric-onset inflammatory bowel disease. *Nat. Genet.* 40:1211–15
78. Laitinen T, Polvi A, Rydman P, Vendelin J, Pulkkinen V, et al. 2004. Characterization of a common susceptibility locus for asthma-related traits. *Science* 304:300–4
79. Larsen JM, Musavian HS, Butt TM, Ingvorsen C, Thysen AH, Brix S. 2015. Chronic obstructive pulmonary disease and asthma-associated Proteobacteria, but not commensal *Prevotella* spp., promote Toll-like receptor 2-independent lung inflammation and pathology. *Immunology* 144:333–42
80. Lasky-Su J, Himes BE, Raby BA, Klanderman BJ, Sylvia JS, et al. 2012. HLA-DQ strikes again: genome-wide association study further confirms HLA-DQ in the diagnosis of asthma among adults. *Clin. Exp. Allergy* 42:1724–33
81. Lauc G, Huffman JE, Pucic M, Zgaga L, Adamczyk B, et al. 2013. Loci associated with N-glycosylation of human immunoglobulin G show pleiotropy with autoimmune diseases and haematological cancers. *PLOS Genet*. 9:e1003225
82. Li Q, Wineinger NE, Fu DJ, Libiger O, Alphs L, et al. 2017. Genome-wide association study of paliperidone efficacy. *Pharmacogenet. Genom.* 27:7–18
83. Li QS, Tian C, Seabrook GR, Drevets WC, Narayan VA. 2016. Analysis of 23andMe antidepressant efficacy survey data: implication of circadian rhythm and neuroplasticity in bupropion response. *Transl. Psychiatry* 6:e889
84. Liang L, Cookson WOC. 2014. Grasping nettles: cellular heterogeneity and other confounders in epigenome-wide association studies. *Hum. Mol. Genet.* 23:R83–88

Downloaded from www.annualreviews.org. Guest (guest) IP: 24.117.150.43 On: Tue, 14 Oct 2025 03:05:08

85. Liang L, Willis-Owen SAG, Laprise C, Wong KC, Davies GA, et al. 2015. An epigenome-wide association study of total serum immunoglobulin E concentration. *Nature* 520:670–74
86. Lichtenstein P, Svartengren M. 1997. Genes, environments, and sex: factors of importance in atopic diseases in 7–9-year-old Swedish twins. *Allergy* 52:1079–86
87. Liu JZ, van Sommeren S, Huang H, Ng SC, Alberts R, et al. 2015. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* 47:979–86
88. Liu L, Pan Y, Zhu Y, Song Y, Su X, et al. 2017. Association between rhinovirus wheezing illness and the development of childhood asthma: a meta-analysis. *BMJ Open* 7:e013034
89. Liu YP, Rajamanikham V, Baron M, Patel S, Mathur SK, et al. 2017. Association of ORMDL3 with rhinovirus-induced endoplasmic reticulum stress and type I Interferon responses in human leucocytes. *Clin. Exp. Allergy* 47:371–82
90. Lluis A, Schedel M, Liu J, Illi S, Depner M, et al. 2011. Asthma-associated polymorphisms in 17q21 influence cord blood *ORMDL3* and *GSDMA* gene expression and IL-17 secretion. *J. Allergy Clin. Immunol.* 127:1587–94
91. Loffredo LF, Abdala-Valencia H, Anekalla KR, Cuervo-Pardo L, Gottardi CJ, Berdnikovs S. 2017. Beyond epithelial-to-mesenchymal transition: common suppression of differentiation programs underlies epithelial barrier dysfunction in mild, moderate, and severe asthma. *Allergy* 72:1988–2004
92. Löser S, Gregory LG, Zhang Y, Schaefer K, Walker SA, et al. 2017. Pulmonary ORMDL3 is critical for induction of *Alternaria*-induced allergic airways disease. *J. Allergy Clin. Immunol.* 139:1496–507
93. Lugogo NL, Kraft M, Dixon AE. 2010. Does obesity produce a distinct asthma phenotype? *J. Appl. Physiol.* 108:729–34
94. Lutz SM, Cho MH, Young K, Hersh CP, Castaldi PJ, et al. 2015. A genome-wide association study identifies risk loci for spirometric measures among smokers of European and African ancestry. *BMC Genet.* 16:138
95. Man WH, de Steenhuijsen Piters WA, Bogaert D. 2017. The microbiota of the respiratory tract: gatekeeper to respiratory health. *Nat. Rev. Microbiol.* 15:259–70
96. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, et al. 2009. Finding the missing heritability of complex diseases. *Nature* 461:747–53
97. Marenholz I, Esparza-Gordillo J, Ruschendorf F, Bauerfeind A, Strachan DP, et al. 2015. Meta-analysis identifies seven susceptibility loci involved in the atopic march. *Nat. Commun.* 6:8804
98. Marsh RL, Kaestli M, Chang AB, Binks MJ, Pope CE, et al. 2016. The microbiota in bronchoalveolar lavage from young children with chronic lung disease includes taxa present in both the oropharynx and nasopharynx. *Microbiome* 4:37
99. Martin EM, Clapp PW, Rebuli ME, Pawlak EA, Glista-Baker E, et al. 2016. E-cigarette use results in suppression of immune and inflammatory-response genes in nasal epithelial cells similar to cigarette smoke. *Am. J. Physiol. Lung Cell. Mol. Physiol.* 311:L135–44
100. Martinez FD, Graves PE, Baldini M, Solomon S, Erickson R. 1997. Association between genetic polymorphisms of the β_2-adrenoceptor and response to albuterol in children with and without a history of wheezing. *J. Clin. Investig.* 100:3184–88
101. McClenaghan J, Warrington NM, Jamrozik EF, Hui J, Beilby JP, et al. 2009. The *PHF11* gene is not associated with asthma or asthma phenotypes in two independent populations. *Thorax* 64:620–25
102. McGovern DP, Gardet A, Torkvist L, Goyette P, Essers J, et al. 2010. Genome-wide association identifies multiple ulcerative colitis susceptibility loci. *Nat. Genet.* 42:332–37
103. McKay JD, Hung RJ, Han Y, Zong X, Carreras-Torres R, et al. 2017. Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. *Nat. Genet.* 49:1126–32
104. Melén E, Granell R, Kogevinas M, Strachan D, Gonzalez JR, et al. 2013. Genome-wide association study of body mass index in 23 000 individuals with and without asthma. *Clin. Exp. Allergy* 43:463–74
105. Melén E, Himes BE, Brehm JM, Boutaoui N, Klanderman BJ, et al. 2010. Analyses of shared genetic factors between asthma and obesity in children. *J. Allergy Clin. Immunol.* 126:631–37
106. Melén E, Kho AT, Sharma S, Gaedigk R, Leeder JS, et al. 2011. Expression analysis of asthma candidate genes during human and murine lung development. *Respir. Res.* 12:86

Downloaded from www.annualreviews.org. Guest (guest) IP: 24.117.150.43 On: Tue, 14 Oct 2025 03:05:08

107. Miller M, Rosenthal P, Beppu A, Mueller JL, Hoffman HM, et al. 2014. ORMDL3 transgenic mice have increased airway remodeling and airway responsiveness characteristic of asthma. *J. Immunol.* 192:3475–87
108. Moffatt MF, Gut IG, Demenais F, Strachan DP, Bouzigon E, et al. 2010. A large-scale, consortium-based genomewide association study of asthma. *N. Engl. J. Med.* 363:1211–21
109. Moffatt MF, Kabesch M, Liang L, Dixon AL, Strachan D, et al. 2007. Genetic variants regulating *ORMDL3* expression contribute to the risk of childhood asthma. *Nature* 448:470–73
110. Newton C, McHugh S, Widen R, Nakachi N, Klein T, Friedman H. 2000. Induction of interleukin-4 (IL-4) by legionella pneumophila infection in BALB/c mice and regulation of tumor necrosis factor α, IL-6, and IL-1β. *Infect. Immun.* 68:5234–40
111. Nicodemus-Johnson J, Myers RA, Sakabe NJ, Sobreira DR, Hogarth DK, et al. 2016. DNA methylation in lung cells is associated with asthma endotypes and genetic risk. *JCI Insight* 1:e90151
112. Nicolae D, Cox NJ, Lester LA, Schneider D, Tan Z, et al. 2005. Fine mapping and positional candidate studies identify *HLA-G* as an asthma susceptibility gene on chromosome 6p21. *Am. J. Hum. Genet.* 76:349–57
113. Noguchi E, Sakamoto H, Hirota T, Ochiai K, Imoto Y, et al. 2011. Genome-wide association study identifies *HLA-DP* as a susceptibility gene for pediatric asthma in Asian populations. *PLOS Genet*. 7:e1002170
114. Noguchi E, Yokouchi Y, Zhang J, Shibuya K, Shibuya A, et al. 2005. Positional identification of an asthma susceptibility gene on human chromosome 5q33. *Am. J. Respir. Crit. Care Med.* 172:183–88
115. Ober C, Hoffjan S. 2006. Asthma genetics 2006: the long and winding road to gene discovery. *Genes Immun*. 7:95–100
116. Ober C, Yao TC. 2011. The genetics of asthma and allergic disease: a 21st century perspective. *Immunol. Rev.* 242:10–30
117. Okbay A, Baselmans BM, De Neve JE, Turley P, Nivard MG, et al. 2016. Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. *Nat. Genet.* 48:624–33
118. Palmer CN, Irvine AD, Terron-Kwiatkowski A, Zhao Y, Liao H, et al. 2006. Common loss-of-function variants of the epidermal barrier protein filaggrin are a major predisposing factor for atopic dermatitis. *Nat. Genet.* 38:441–46
119. Parsa A, Chang YP, Kelly RJ, Corretti MC, Ryan KA, et al. 2011. Hypertrophy-associated polymorphisms ascertained in a founder cohort applied to heart failure risk and mortality. *Clin. Transl. Sci.* 4:17–23
120. Patarcic I, Gelemanovic A, Kirin M, Kolcic I, Theodoratou E, et al. 2015. The role of host genetic factors in respiratory tract infectious diseases: systematic review, meta-analyses and field synopsis. *Sci. Rep.* 5:16119
121. Paternoster L, Standl M, Waage J, Baurecht H, Hotze M, et al. 2015. Multi-ancestry genome-wide association study of 21,000 cases and 95,000 controls identifies new risk loci for atopic dermatitis. *Nat. Genet.* 47:1449–56
122. Peters-Golden M, Swern A, Bird SS, Hustad CM, Grant E, Edelman JM. 2006. Influence of body mass index on the response to asthma controller agents. *Eur. Respir. J.* 27:495–503
123. Pickrell JK, Berisa T, Liu JZ, Segurel L, Tung JY, Hinds DA. 2016. Detection and interpretation of shared genetic influences on 42 human traits. *Nat. Genet.* 48:709–17
124. Ramasamy A, Kuokkanen M, Vedantam S, Gajdos ZK, Couto Alves A, et al. 2012. Genome-wide association studies of asthma in population-based cohorts confirm known and suggested loci and identify an additional association near HLA. *PLOS ONE* 7:e44008
125. Rastogi D, Fraser S, Oh J, Huber AM, Schulman Y, et al. 2015. Inflammation, metabolic dysregulation, and pulmonary function among obese urban adolescents with asthma. *Am. J. Respir. Crit. Care Med.* 191:149–60
126. Rosenberg HF, Dyer KD, Foster PS. 2013. Eosinophils: changing perspectives in health and disease. *Nat. Rev. Immunol.* 13:9–22
127. Shindou H, Hishikawa D, Nakanishi H, Harayama T, Ishii S, et al. 2007. A single enzyme catalyzes both platelet-activating factor production and membrane biogenesis of inflammatory cells. Cloning and characterization of acetyl-CoA:LYSO-PAF acetyltransferase. *J. Biol. Chem.* 282:6532–39

Downloaded from www.annualreviews.org. Guest (guest) IP: 24.117.150.43 On: Tue, 14 Oct 2025 03:05:08

128. Silverberg MS, Cho JH, Rioux JD, McGovern DP, Wu J, et al. 2009. Ulcerative colitis-risk loci on chromosomes 1p36 and 12q15 found by genome-wide association study. *Nat. Genet.* 41:216–20
129. Sleiman PM, Flory J, Imielinski M, Bradfield JP, Annaiah K, et al. 2010. Variants of *DENND1B* associated with asthma in children. *N. Engl. J. Med.* 362:36–44
130. Smit LA, Bouzigon E, Pin I, Siroux V, Monier F, et al. 2010. 17q21 variants modify the association between early respiratory infections and asthma. *Eur. Respir. J.* 36:57–64
131. Song DJ, Miller M, Beppu A, Rosenthal P, Das S, et al. 2017. Rhinovirus infection of ORMDL3 transgenic mice is associated with reduced rhinovirus viral load and airway inflammation. *J. Immunol.* 199:2215–24
132. Song Y, Schwager MJ, Backer V, Guo J, Porsbjerg C, et al. 2017. Environment changes genetic effects on respiratory conditions and allergic phenotypes. *Sci. Rep.* 7:6342
133. Soriano JB, Davis KJ, Coleman B, Visick G, Mannino D, Pride NB. 2003. The proportional Venn diagram of obstructive lung disease: two approximations from the United States and the United Kingdom. *Chest* 124:474–81
134. Spada J, Scholz M, Kirsten H, Hensch T, Horn K, et al. 2016. Genome-wide association analysis of actigraphic sleep phenotypes in the LIFE Adult Study. *J. Sleep Res.* 25:690–701
135. Staley JT, Konopka A. 1985. Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. *Annu. Rev. Microbiol.* 39:321–46
136. Strachan DP. 1989. Hay fever, hygiene, and household size. *BMJ* 299:1259–60
137. Torgerson DG, Ampleford EJ, Chiu GY, Gauderman WJ, Gignoux CR, et al. 2011. Meta-analysis of genome-wide association studies of asthma in ethnically diverse North American populations. *Nat. Genet.* 43:887–92
138. UK IBD Genet. Consort., Wellcome Trust Case Control Consort. 2009. Genome-wide association study of ulcerative colitis identifies three new susceptibility loci, including the *HNF4A* region. *Nat. Genet.* 41:1330–34
139. Umland SP, Garlisi CG, Shah H, Wan Y, Zou J, et al. 2003. Human *ADAM33* messenger RNA expression profile and post-transcriptional regulation. *Am. J. Respir. Cell Mol. Biol.* 29:571–82
140. van Beijsterveldt CE, Boomsma DI. 2007. Genetics of parentally reported asthma, eczema and rhinitis in 5-yr-old twins. *Eur. Respir. J.* 29:516–21
141. van der Valk RJ, Duijts L, Kerkhof M, Willemsen SP, Hofman A, et al. 2012. Interaction of a 17q12 variant with both fetal and infant smoke exposure in the development of childhood asthma-like symptoms. *Allergy* 67:767–74
142. Van Eerdewegh P, Little RD, Dupuis J, Del Mastro RG, Falls K, et al. 2002. Association of the *ADAM33* gene with asthma and bronchial hyperresponsiveness. *Nature* 418:426–30
143. Vonk JM, Scholtens S, Postma DS, Moffatt MF, Jarvis D, et al. 2017. Adult onset asthma and interaction between genes and active tobacco smoking: the GABRIEL consortium. *PLOS ONE* 12:e0172716
144. Wain LV, Shrine NR, Soler Artigas M, Erzurumluoglu AM, Noyvert B, et al. 2017. Genome-wide association analyses for lung function and chronic obstructive pulmonary disease identify new loci and potential druggable targets. *Nat. Genet.* 49:416–25
145. Wan YI, Shrine NR, Soler Artigas M, Wain LV, Blakey JD, et al. 2012. Genome-wide association study to identify genetic determinants of severe asthma. *Thorax* 67:762–68
146. Wang K, Baldassano R, Zhang H, Qu HQ, Imielinski M, et al. 2010. Comparative genetic analysis of inflammatory bowel disease and type 1 diabetes implicates multiple loci with opposite effects. *Hum. Mol. Genet.* 19:2059–67
147. Wang L, Murk W, DeWan AT. 2015. Genome-wide gene by environment interaction analysis identifies common SNPs at 17q21.2 that are associated with increased body mass index only among asthmatics. *PLOS ONE* 10:e0144114
148. Weidinger S, Willis-Owen SAG, Kamatani Y, Baurecht H, Morar N, et al. 2013. A genome-wide association study of atopic dermatitis identifies loci with overlapping effects on asthma and psoriasis. *Hum. Mol. Genet.* 22:4841–56
149. Weinmayr G, Forastiere F, Buchele G, Jaensch A, Strachan DP, Nagel G. 2014. Overweight/obesity and respiratory and allergic disease in children: international study of asthma and allergies in childhood (ISAAC) phase two. *PLOS ONE* 9:e113996

Downloaded from www.annualreviews.org. Guest (guest) IP: 24.117.150.43 On: Tue, 14 Oct 2025 03:05:08

150. Weiss ST. 2005. Obesity: insight into the origins of asthma. *Nat. Immunol.* 6:537–39
151. White JH, Chiano M, Wigglesworth M, Geske R, Riley J, et al. 2008. Identification of a novel asthma susceptibility gene on chromosome 1qter and its functional evaluation. *Hum. Mol. Genet.* 17:1890–903
152. World Health Organ. (WHO). 1990. *International Statistical Classification of Diseases and Related Health Problems*. Geneva: WHO. 10th rev.
153. Xu C-J, Söderhäll C, Bustamante M, Baïz N, Gruzieva O, et al. 2018. DNA methylation in childhood asthma: an epigenome-wide meta-analysis. *Lancet Respir. Med.* 6:379–88
154. Xu W, Xu J, Shestopaloff K, Dicks E, Green J, et al. 2015. A genome wide association study on Newfoundland colorectal cancer patients' survival outcomes. *Biomark. Res.* 3:6
155. Yamazaki K, Umeno J, Takahashi A, Hirano A, Johnson TA, et al. 2013. A genome-wide association study identifies 2 susceptibility loci for Crohn's disease in a Japanese population. *Gastroenterology* 144:781–88
156. Yan Q, Brehm J, Pino-Yanes M, Forno E, Lin J, et al. 2017. A meta-analysis of genome-wide association studies of asthma in Puerto Ricans. *Eur. Respir. J.* 49:1601505
157. Yan X, Chu JH, Gomez J, Koenigs M, Holm C, et al. 2015. Noninvasive analysis of the sputum transcriptome discriminates clinical phenotypes of asthma. *Am. J. Respir. Crit. Care Med.* 191:1116–25
158. Yang SK, Hong M, Zhao W, Jung Y, Baek J, et al. 2014. Genome-wide association study of Crohn's disease in Koreans revealed three new susceptibility loci and common attributes of genetic susceptibility across ethnic populations. *Gut* 63:80–87
159. Yang SK, Hong M, Zhao W, Jung Y, Tayebi N, et al. 2013. Genome-wide association study of ulcerative colitis in Koreans suggests extensive overlapping of genetic susceptibility with Caucasians. *Inflamm. Bowel Dis.* 19:954–66
160. Zeng Z, Feingold E, Wang X, Weeks DE, Lee M, et al. 2014. Genome-wide association study of primary dentition pit-and-fissure and smooth surface caries. *Caries Res.* 48:330–38
161. Zhang Y, Fear DJ, Willis-Owen SAG, Cookson WOC, Moffatt MF. 2016. Global gene regulation during activation of immunoglobulin class switching in human B cells. *Sci. Rep.* 6:37988
162. Zhang Y, Leaves NI, Anderson GG, Ponting CP, Broxholme J, et al. 2003. Positional cloning of a quantitative trait locus on chromosome 13q14 that influences immunoglobulin E levels and asthma. *Nat. Genet.* 34:181–86

Downloaded from www.annualreviews.org. Guest (guest) IP: 24.117.150.43 On: Tue, 14 Oct 2025 03:05:08

Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

HHS Public Access
Author manuscript
Curr Cardiol Rep. Author manuscript; available in PMC 2017 June 01.

Published in final edited form as:
Curr Cardiol Rep. 2016 June ; 18(6): 55. doi:10.1007/s11886-016-0735-8.

Genomics of Atrial Fibrillation

Alejandra Gutierrez, MD[b] and **Mina K. Chung, MD**[a]

[a]Department of Cardiovascular Medicine, Cleveland Clinic, Cleveland, OH; Department of Molecular Cardiology, Cleveland Clinic, Cleveland, OH

[b]Department of Internal Medicine, Cleveland Clinic, Cleveland, OH

Abstract

Atrial fibrillation (AF) is a common clinical arrhythmia that appears to be highly heritable, despite representing a complex interplay of several disease processes that generally do not manifest until later in life. In this manuscript we will review the genetic basis of this complex trait established through studies of familial AF, linkage and candidate gene studies of common AF, genome wide association studies (GWAS) of common AF, and transcriptomic studies of AF. Since AF is associated with a five-fold increase in the risk of stroke, we also review the intersection of common genetic factors associated with both of these conditions. Similarly, we highlight the intersection of common genetic markers associated with some risk factors for AF, such as hypertension and obesity, and AF. Lastly, we describe a paradigm where genetic factors predispose to the risk of AF, but which may require additional stress and trigger factors in older age to allow for the clinical manifestation of AF.

Keywords

Atrial Fibrillation; Genomics; Stroke; Hypertension

Introduction

Atrial fibrillation (AF) is the most common sustained arrhythmia seen in clinical practice. The etiology of AF is multifactorial, but includes a genetic predisposition to the arrhythmia. The road towards discovering the genetic basis of AF has progressed from identification of genes associated with familial AF, to clinical observational studies demonstrating hereditability of common AF, to genome wide association studies (GWAS) that have identified to date 14 genetic loci associated with AF (Table 1)[1–16]. However, identification of risk loci is only the start of a long process to discover the mechanisms by which these variants increase AF risk. Moreover, despite the identification of inherent

---

Corresponding Author: Mina K. Chung, MD, Department of Cardiovascular Medicine, Heart & Vascular Institute, Department of Molecular Cardiology, Lerner Research Institute, Cleveland Clinic, 9500 Euclid Ave., J2-2, Cleveland, OH 44195, Office: 216-444-2290, FAX: 216-636-6951, chungm@ccf.org.

Conflict of Interest
Alejandra Gutierrez and Mina K. Chung declare that they have no conflict of interest.

Human and Animal Rights and Informed Consent
This article does not contain any studies with human or animal subjects performed by any of the authors.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

genetic risk, we do not fully understand why AF typically does not manifest until later decades of life, nor why certain clinical factors, such as hypertension, obesity, male sex, or Caucasian ancestry, predispose to AF. Here, we will review the published evidence for the genetic basis of AF and discuss the results of recent genomic and transcriptomic analyses that may yield insights into some clinical observations of AF incidence.

Heritability of Atrial Fibrillation

The lifetime risk for developing AF is one in four for people older than 40 years [17]. Traditional risk factors for AF include older age, obesity, alcohol consumption, hypertension, left ventricular hypertrophy, heart failure, prolonged PR interval and coronary artery disease [18]. The prevalence of AF has increased as a result of the aging and increased prevalence of obesity in the general population and in response to an increased awareness and diagnosis of the arrhythmia [18].

Despite the complex and aging-related clinical risk factors for the disease, AF is heritable. In the Framingham study parental AF increased the risk of AF in offspring [19]. The odds ratio was 1.85 for AF in offspring who had one parent with AF and 3.23 when the sample was restricted to patients with parents who developed AF when they were <75years old. A more recent study involving the Framingham cohort reported that new onset AF occurred more frequently in subjects with compared to without familial AF (adjusted hazard ratio 1.40; 95% CI 1.13–1.74), independent of traditional AF risk factors [20]. In Iceland, Arnar et al. studied 5269 patients diagnosed with AF and age- and sex-matched controls[1]. The relative risk of AF in first degree relative pairs was 1.77. When the analysis was restricted to first degree relatives of patients with early onset AF (diagnosed <60 years of age), relative risk of AF was 4.67. Ellinor and colleagues reported that lone AF (i.e. AF in the absence of structural heart disease) greatly increases the risk of AF in family members. A family history of AF was found in 38% of patients with lone AF; the relative risk for sons was 8.1, daughters 9.5, brothers 70, and sisters 34[21]. Lastly, a Danish twin study reported 22.0% concordance rates for monozygotic pairs compared to 11.6% for dizygotic pairs ($p<0.0001$) with heritability of AF estimated to be 62%[22]. Of note, such twin studies are of particular interest because they are less likely to lead to inflated estimates of heritability as a consequence of shared environmental risk factors.

Familial Monogenic AF

Families exhibiting a very high prevalence of AF have been reported as early as 1936 [23–25]. Traditional approaches to identifying genes responsible for AF in these presumably monogenic (or Mendelian) forms of AF include linkage studies and candidate gene searches. These studies have identified several ion channel gene mutations that were associated with familial AF, including potassium channel mutations (e.g. in *KCNQ1, KCNE2, KCNH2, KCNJ2, KCNA5*) for which gain-of-function mutations might be expected to promote AF by shortening of action potential duration and effective refractory periods, as well as in the sodium channel mutations (in *SCN5A* and *SCN4B*), which have also been associated with dilated cardiomyopathy[26–35].

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Several non-ion channel genes have also been associated with familial monogenic AF, including *PRKAG2*, associated with familial Wolff-Parkinson-White syndrome and AF[36], and *NPPA*, which encodes for atrial natriuretic peptide [37]. Further, some developmental genes have been implicated, including genes encoding transcription factors involved in cardiogenesis from the *GATA* family[38–40] and *NKX2-5*, a homeobox-containing transcription factor involved in cardiac development and septation with mutations associated with atrial septal defect and AV conduction abnormalities [41]. Structural genes have also been associated with familial AF, including *GJA5*, which encodes for connexin 40 [42], a gap junction protein that is expressed in the atria and the conduction system and that is crucial for action potential propagation through myocytes, and *RYR2*, a gene which encodes the ryanodine receptor responsible for releasing intracellular stores of calcium [43, 44]. Finally, genes encoding nuclear membrane associated proteins have been linked to familial AF. These include *LMNA*, which encodes for lamin expressed in the inner nuclear lamina. Mutations in *LMNA* have also been associated with muscular dystrophy, dilated cardiomyopathy, risk of sudden cardiac death[2, 45], and progeria syndrome. *NUP155* has also been associated with AF[46, 47]. This gene encodes for a nucleoporin that is involved in the assembly and functioning of the nuclear pore complex that regulates transport of macromolecules across the nuclear envelope between the cytoplasm and the nucleus [46, 47].

Polygenic Non-familial Common AF

Genetic Architecture

Although familial AF studies have contributed much to our understanding of the genes associated with monogenic forms of AF, they represent only a small fraction of the overall burden of disease. Overall, the genetics of common AF pathogenesis is complex, involving modest contributions to the risk of disease from genetic variations in many genes. The genetic architecture of a disease depends on the number of variants that influence a given pathology, the effect size of each variant on the phenotype, and the frequency of the variant. Complex diseases such as AF follow the "common disease common variant" hypothesis of genetic architecture where common variants with small effects appear to be responsible for a majority of the heritability of the trait[48].

Linkage and Candidate Gene Studies

Initial strategies to identify polymorphisms predisposing to common polygenic AF included linkage analyses and candidate gene based association studies. Linkage analyses attempted to apply the same analytic techniques used to identify monogenic forms of disease to common polygenic forms of disease but proved to be inadequately powered to detect the modest effects of polymorphisms contributing to the risk of common AF [49].

Failed linkage studies were followed by candidate gene association studies, which relied on prior knowledge to generate hypotheses about possible polymorphisms that could have a causal relationship with a given disease and were limited by the potential bias of studying only what is already known [50]. Since these studies concentrated on only a small region of the genome, their pretest probability of identifying a real signal was low. The data published

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

was further influenced by publication bias. A meta-analysis of 370 studies reporting on 36 different genetic associations illustrated this point well. The study found that in 25 of 36 cases, the first study overestimated the replicated effect[49]. Further, 39% of the genetic associations studied had significant heterogeneity between results [49]. Lack of replication for candidate gene association studies is common and can be a result of differences in study design, differences in studied populations with variable allele frequencies or relative risks of a certain variant, or due to real biological differences[50]. Furthermore, the polygenic nature of common and complex diseases involving many variants with small relative risks resulted in many studies being substantially underpowered leading to a high rate of false positive reports [49].

While as a whole candidate gene association studies have been disappointing, such studies nevertheless yielded some valuable insights into the potential mechanisms leading to AF. To begin, such studies implicated ion channel genes as contributants to risk of common AF, including those identified in familial studies. Potassium channel genes, including *KCNQ1, KCNH2, KCNJ5, KCNJ8, KCND3* and *KCNE 1* through *5* have been associated with AF risk[13, 51–57]. The *KCNE* genes encode the beta subunits of voltage gated channels; *KCNE1* and *KCNQ1* contribute to the I_{Ks} current [52–54, 58]. Gain of function of I_{Ks} [52, 59] results in a faster repolarizing current and shorter effective refractory period, increasing the excitability of cells and susceptibility to arrhythmia. Loss of function prolongs the action potential and increases the likelihood of early after depolarizations [60]. *KCNA5* encodes for the α-subunit of the voltage sensor of a channel that generates the ultra-rapid potassium current, I_{Kur}, which is abundant in atria and detectable in failing ventricles. Hayashi et al. found a novel gain of function variant and a previously described *KCNA5* loss of function variant in a cohort of 72 patients with lone AF [13, 61].

Mutations in sodium channel subunit genes have also been associated with AF. The voltage gated sodium channel consists of a pore forming α subunit, encoded by the *SCNA* genes, and 4 different β subunits encoded by *SCN1B-4B*. This channel is responsible for the rapid influx of sodium ions that underlies the upstroke (phase 0) of the cardiac action potential. Several variants of *SCNA5*, which encodes the cardiac α-subunit of Nav1.5, have been reported in patients with common polygenic AF [62]. Gain of function mutations decrease the threshold voltage for the action potential and increase cellular excitability[63]. β-subunits modulate the gating and voltage dependence of the channels, impact the trafficking of sodium channel to and from the cell membrane, and aid in cell adhesion. Mutations in the four β-subunits have been reported to attenuate sodium currents and shift the voltage-dependence of sodium channel gating [64, 65]. A single nucleotide polymorphism (SNP) in *SCN10A*, which encodes Nav1.8, a voltage gated sodium channel that participates in the late sodium current, was recently described to be associated with early onset AF [66].

Candidate gene analyses have also identified polymorphisms in non-ion channel genes with familial AF. Implicated genes include those encoding connexin 40, *GJA5* [7, 67–69], the transcription factor gene *NKX2.5*[70, 71], and the *LMNA* gene[72]. Genes involved in the renin-angiotensin-aldosterone pathway, including the angiotensin converting enzyme, angiotensin gene promoter and angiotensinogen[73–76], have also been associated with AF.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Identification of AF-associated Genetic Loci Through Genome Wide Association Studies

Genome wide association studies (GWAS) represent a relatively new but important study design to overcome the inherent bias of candidate gene analyses. By comparing the frequencies of common genetic variants among subjects with and without a disease or level of quantitative trait, GWAS has been able to identify hundreds of novel genetic associations over the last 10 years involving common single nucleotide polymorphisms (SNPs). GWAS analyze hundreds of thousands to millions of SNPs concurrently. To correct for multiple testing, thresholds for significance typically need to be adjusted to a p value of $< 5\times10^{-8}$, which effectively corrects for 1 million independent tests. Once a SNP or a set of SNPs in a region reaches genome wide significance, the next important step is to link the SNP association to a causal gene. Generally, this process begins with an assessment of the genes that are in proximity or within the susceptibility locus and should not be deemed certain until a functional mechanism has been fully worked out in detail through multiple additional experiments[77].

Developmental Genes

In 2007, the first GWAS for AF identified a susceptibility locus at chromosome 4q25 in a population of European descent and replicated this association in an Asian population [12]. This association has been further replicated in all AF GWAS to date[78–80]. Further, this locus was associated with recurrence of AF after ablation and new onset AF in the postoperative period after coronary bypass surgery [3, 78].

Additional GWAS have identified at least 4 independent loci at chromosome 4q25 for AF [81]. As is often the case, the lead SNPs in this region are in non-coding regions. Nevertheless, the closest gene is the paired-like homeodomain 2 *PITX2* gene, located in a linkage disequilibrium (LD) block upstream to the susceptibility locus. *PITX2* is involved in embryogenesis and left-right differentiation of the heart. *PITX2* encodes 3 different isoforms of the protein by alternative splicing. Importantly, the Pitx2c isoform is mainly found in the left atrium and pulmonary veins [82][83]. In Pitx2c conditional knockout mice, Pitx2c-deficient mice do not develop a pulmonary myocardial sleeve, and Pitx2 appears to suppress a left atrial sinus node program [84]. Compared to wild type mice, Pitx2null+/− and Pitx2null−/− mice heterozygous and homozygous for a Pitx2-null allele that removes all Pitx2 isoform function showed step-wise increased expression of *HCN4*, which encodes a hyperpolarization-activated cyclic nucleotide-gated channel contributing to pacemaker currents, as well as *TBX*3 and *SHOX2* that are also required for sinoatrial node development[83]. Expression of *KCNQ1* was also increased[84], which would result in a shorter effective refractory period, facilitating action potential triggering and conduction. Mouse *PITX2* conditional or heterozgygote mutation models have shown alterations in ion channel, calcium handling, gap and tight junction gene expression[85–88], *PITX2* regulation of miRNAs[87–89], and atrial arrhythmia inducibility [84, 85, 87]. Thus, *PITX2* is a biologically attractive candidate causative gene in the region, as it appears to be involved in the regulation of several ion channel genes that plausibly may predispose to atrial arrhythmia, it directs asymmetric morphogenesis of the heart, is involved in the formation of

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

pulmonary veins, the putative site of AF initiation targeted by clinical AF ablation [90], and appears to repress a left atrial sinus node program, genetic variation of which might contribute to electrical triggers recorded within pulmonary vein ostia.

The expression of *PITX2* has been shown to markedly decrease postnatally. The mechanisms leading to postnatal inactivation of *PITX2* are actively being researched and include possible mediation by various risk factors, such as pressure and volume overload and oxidative stress that come with age. Tao, et al, showed that postnatal inactivation of *PITX2* leads to abnormal R-R intervals and to differential expression of genes involved in cell junction organization, including the intercalated disk, ion channels and caveolae. [91]

GWAS of AF have also identified several other loci associated with transcription factors that appear to be important for cardiac development. One of these is located on chromosome 16q22 intronic to the zinc finger homeobox 3 gene (*ZFHX3*), known as AT motif-binding transcription factor 1(*ATBF1*) [4], which is involved in regulation of growth and differentiation of skeletal muscle and neuronal tissues [4]. *ZFHX3* is required for the transcriptional activation of the gene POU class I homebox 1 (*POU1F1*), which interacts with *PITX2*, enabling DNA binding and transcriptional activity. This might be a mechanism by which it predisposes to AF [80]. Later, another GWAS described *PRRX1* which encodes for a homeodomain transcription factor expressed in connective tissue in the developing heart[10]. In animal models, knockout of *PRRX1* led to abnormalities in the pulmonary vasculature [92].

Ion Channels and Pores

Thus far, two loci in ion channel genes have been significantly associated with AF in GWAS studies. First, a locus was identified on 1q22 within an intron of the *KCNN3* gene, which encodes for the calcium activated potassium channel SK3 [11]. SK3 contributes to the repolarization phase of the action potential[93]. Second, a locus on chromosome 15q24 within an intron of the hyperpolarization-activated cyclic nucleotide-gated 4 (*HCN4*) gene was described [10]. *HCN4* is expressed in the sinoatrial (SA) node and underlies the I_f current normally responsible for the pacemaker current in nodal myocytes. Mutations of this channel may lead to diminished action potential frequency (heart rate slowing) and delayed after depolarizations that might trigger AF [94]. As discussed above, Pitx2 represses Hcn4 expression in the left atrium[84][83].

Structural Genes

A meta-analysis of GWAS in 2010 identified several new AF loci, including three in/near genes involved in cellular structural components. Cell membranes contain invaginations, known as caveolae, that participate in cell transport. A locus on chromosome 7q31 within an intron of the gene *CAV1*, which encodes caveolin-1, was associated with AF[10]. *CAV1* is co-expressed with a potassium channel gene (*KCNH2*), which is involved in the repolarization phase of the action potential. The risk allele and AF are associated with decreased expression of *CAV1*[95]. Further caveolin-1 has been postulated as an anti-fibrotic element in atrial tissue, since down-regulation of Cav-1 by siRNA leads to increased TGF-β1-induced activation of the Smad signal pathway and collagen production[96].

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

An intronic SNP located on chromosome 14q23 in the gene *SYNE2*, which encodes for nesprin-2, has also been associated with AF. Nesprin-2 is located in the outer nuclear membrane and sarcomere and anchors nuclei to the cytoskeleton. A large meta-analysis of GWAS by Ellinor, et al, identified a susceptibility locus between two genes, *SYNPO2L* and *MYOZ1*, on chromosome 10 and another locus in the an open reading frame on chromosome 9 (*C9orf3*) [10]. Myozenin-1 is a protein involved in stabilizing the sarcomere, encoded by *MYOZ1*. *C9orf3* region is associated with three coding genes, including *FBP1* and *FBP2*, which are involved in gluconeogenesis[10].

Gene mRNA Expression studies

Although GWAS studies identify regions in the genome that are associated with a phenotype, the lead SNP within a susceptibility locus is often not the causal variant. Furthermore, the lead SNP is often not within a coding region of a gene and usually near several genes, making it challenging to determine which gene in the region is the causal gene. One way to identify with more certainty the causal gene is to leverage gene expression studies to determine whether a lead SNP is preferentially associated with the level of mRNA of one or more genes nearby. SNPs that are associated with mRNA are often termed expression quantitative trait loci (eQTLs) [97]. As the control of gene expression is often tissue-specific, assessment of relationships between SNPs and mRNA expression among circulating cells in blood may be inadequate to assess the effects of SNPs on the expression of genes predisposing to AF. Gene expression studies involving more relevant target tissues, such as the left atrium itself, may substantially increase the yield of identifying causal genes among susceptibility loci of AF.

Genetic variants can change the expression of mRNA and proteins by *cis* or *trans* mechanisms. *Cis*-acting variants affect expression of genes nearby on the same chromosome as the gene variants, while *trans*-acting variants can act on genes that are far from them. Such transcriptional studies can help guide the identification of the genes associated with genetic variants identified by GWAS.

eQTL analysis of the AF-associated locus in between *SYNPO2L* and *MYOZ1* has demonstrated an association between the high risk alleles in this region and decreased expression of *MYOZ1*, implicating *MYOZ1* as the most likely causal gene associated with this AF locus [95, 98]. In another eQTL study predominantly involving right atrial appendage tissues, AF risk alleles were associated with increased expression of *PITX2a* and decreased expression of *MYOZ1, CAV1, C9orf3*, and *FANCC* [95].

Despite the finding of cis-eQTLs for genes noted above, significant eQTLs for the top GWAS locus from AF on chromosome 4q25 were not found for *PITX2c* in human adult left atrial appendage tissues [99]. Recently, we performed RNA sequencing on human left/right atrial tissue pairs and reported the presence of an intergenic long non-coding RNA sequence adjacent to the *PITX2* gene (PANCR) differentially expressed in the left atria and eye but no other heart chambers. Expression of PANCR RNA correlated with *PITX2c* mRNA expression. In human embryonic stem cells, both PANCR and *PITX2c* RNAs were coordinately expressed during early differentiation of the stem cells towards

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

cardiomyocytes. siRNA-mediated PANCR knockdown decreased *PITX2c* expression and resulted in similar RNA expression profiles as *PITX2c* knockdown cells, whereas *PITX2c* knockdown did not affect PANCR expression, suggesting that PANCR is an important regulator of *PITX2c* expression [100]. However, similar to *PITX2c* in human adult left atrial appendage tissue, chromosome 4q25 risk SNPs for AF were not associated with PANCR or *PITX2c* expression[99, 100]. These findings may be a consequence of *PITX2* and PANCR being most active during embryologic development in the left atrium or in the tissue within the pulmonary vein – left atrial junction which may have some distinct differences with the tissue making up the adult left atrial appendage.

A combination of large scale GWAS, cis-eQTL and functional evaluation led to the identification of 5 novel loci associated with AF near the genes *NEURL, GJA1, TBX5, CAND2* and *CUX2*[15]. AF-associated cis-eQTLs were identified for *CAND2, GJA1*, and *TBX5*. *NEURL* encodes for an E3 ubiquitin ligase that was found to co-localize with *PITX2*. Another signal was found in a SNP intronic to *TBX5*, a transcription factor implicated in conduction abnormalities, and in *CAND2* [15], which encodes for a TATA binding protein transcription factor involved in myogenesis[101]. eQTL analysis of *CAND2* and *TBX5* showed that the risk alleles were associated with an increased expression of these genes. Additional functional studies demonstrated that *NEURL* and *CAND2* knockout zebrafish models had longer atrial action potential duration[15]. Propagation of the action potential through myocytes requires a balanced expression of two gap junction proteins: connexin 40 and 43 [102]. Cis-eQTL associations showed the AF risk allele at the *GJA1* locus was associated with lower expression of *GJA1*, which encodes connexin 43. [15] Connexin abnormalities can lead to reduced conduction velocity and action potential propagation. In the same paper, a novel SNP located within an intron of the *CUX2* gene was found to be associated with AF only in a Japanese population [15]. *CUX2* was recently associated with increased oxidative DNA damage in embryonic neurons [103]. However, a significant cis-eQTL was not identified for the risk allele.

Transcriptomic Analyses and the Age Paradox in AF

Despite the association of several developmental genes with AF risk, AF typically does not manifest until decades after birth. To gain insight on this paradox, we recently performed a gene expression study of left atrial appendage tissue from 207 adults with AF and 32 adults without AF to compare gene expression in 3 groups based on AF history and rhythm at the time of resection of the atrial appendage: 1) No AF history and in sinus rhythm (SR/SR); 2) AF history and in sinus rhythm (AF/SR); and 3) AF history and in AF rhythm (AF/AF)[9]. We identified differentially coexpressed gene modules between the SR/SR and AF/SR groups (“AF susceptibility” comparison), and between the AF/SR and AF/AF groups (“AF activity or persistence” comparison). Gene set enrichment analyses showed that AF susceptibility was associated with decreased transcriptional responses to cellular stress, inflammation, oxidation, and unfolded proteins, while AF persistence was associated with changes in ion channel expression and decreased stress responses. These findings may help explain the AF age paradox, whereby AF does not manifest until there is impaired responsiveness to persistent cellular stress and aging, decades after the GWAS-identified AF loci exert their activity during cardiac and pulmonary vein development.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

The reduced cellular stress response to protein unfolding is of particular interest, as this geneset includes targets of several transcription factors such as the CREB/ATF family, HSF1, ATF6, SRF, E2F1, and SP1[9]. Alterations in many of these transcription factors can lead to abnormal proteostasis and deposition of unfolded amyloid proteins. Deposition of isolated atrial amyloid is well described and associated with older age and the presence of AF [104, 105]. Interestingly, atrial amyloid has also been reported to be more frequent in left than in right atrial appendages [105]. Isolated atrial amyloid protein has been shown to be due to aggregates of atrial natriuretic peptide (ANP) [104, 106], and expression levels of *NPPA*, which encodes for ANP, are high in adult human atrial appendage tissues. Expression of *NPPA* was up-regulated in Pitx2 heterozygote and null mice [83]. Further, a recent study in a Chinese population suggested that *TBX5* mutations increases the expression of ANP and connexin 40[107]. Senile accumulation of the protein aggregates may contribute to the predisposition to AF with aging.

AF Genetics and Stroke

One of the most devastating complications of AF is stroke. Conventional risk factors account for only 60% of the variation in the risk of stroke [108]. Similar to AF, stroke has a genetic predisposition. Further, risk factors differ depending on the etiology of the stroke, which includes hemorrhagic, ischemic, cardioembolic, large artery due to atherosclerosis, small vessel and cryptogenic forms[109]. Genetic studies have attempted to elucidate new stroke prediction, prevention and treatment approaches and have led to the discovery of common genetic variants associated with both AF and stroke (Table 2, Figure 1).

Soon after publication of the first GWAS of AF implicating SNPs in the 4q25 region, the same SNPs were also identified as risk factors for ischemic stroke. Gretarsdottir, et al, showed that rs2200733, one of the first 4q25 SNPs demonstrated to be strongly associated with AF risk, was associated with ischemic stroke, and even more strongly with cardioembolic stroke [110]. This association was further replicated in a targeted sequencing study of a Polish population[111]. Lastly, another study showed a different locus in the same region was associated only with cardioembolic stroke [112].

Large scale studies have found similar genetic associations between AF and stroke. Gudbjartsson, et al, reported a case control genomic study, finding that the SNP rs7193343 at the *ZFHX3* locus was associated with ischemic and cardioembolic stroke[113]. Several GWAS studies have confirmed this association, including the METASTROKE study, which incorporated about 70,000 subjects. These studies confirmed that the *PITX2* and *ZFHX3* SNPs were associated with ischemic stroke and more specifically with cardioembolic stroke, even though only half of their population was sub classified by type of stroke [114]. A recent exome sequencing study replicated the *ZFHX3* locus association with ischemic stroke[115]. Using the METASTROKE collaboration data, an association between ischemic stroke with an AF-risk SNP intronic to *CUX2* was found, though the AF risk allele was paradoxically associated with a decreased the risk of stroke (OR 0.95.; 95% CI 0.91–0.98). Sub-analysis by type of stroke found an association of *GJA1* and *TBX5* AF-risk SNPs with cardioembolic stroke[15].

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

The mentioned associations are likely due to an enrichment of AF cases among ischemic and cardioembolic stroke cases. They might be markers of AF as a predominant cause of embolic rather than non-embolic stroke. Studies have not consistently classified strokes and the lack of differentiation by subtype of stroke reduces power for stroke subtype analyses. Without a large study of isolated cardioembolic stroke, signals attributed to ischemic stroke are an aggregate of multiple mechanisms leading to stroke. Nevertheless, these genetic variants might ultimately prove useful in differentiating the etiology of an ischemic stroke and in identifying patients with cryptogenic stroke who are most likely to have underlying undiagnosed AF. These patients may warrant more monitoring and potential anticoagulation for secondary prevention of stroke.

Current stroke prevention scores rely on clinical data to determine therapy, but adding genetic markers may improve the predictive value of these tools. Tada, et al, constructed a 12 SNP genetic risk score (including SNPs near *PITX2, KCNN3, PRRX1, CAV1*, C9orf3, *SYNPO2L, SYNE2, HCN4*, *ZFHX3*) to identify people at higher risk for AF and found that the score in addition to the CHADS2 score led to better prediction of ischemic stroke in AF patients[116]. The renin angiotensin aldosterone (RAAS) system plays an important role in inflammation, hypertension, tissue remodeling and fibrosis. A preliminary association of the G6 polymorphism (a polymorphism located in the promoter region of the angiotensin gene leading to increased angiotensinogen production) has been found in stroke in two candidate gene studies [118, 119] and in one candidate gene study of AF patients [76]. Tsai, et al, showed that the addition of the G6 polymorphism to the CHADS2 score improved the ability to predict stroke in an AF population[117]. Other elements of the RAAS system, such as the angiotensinogen gene and angiotensinogen converting enzyme gene, have also been associated with AF and stroke in candidate gene analyses [74, 75, 120, 121].

Inflammation likely plays a role in both stroke and AF. Ischemic brain injury is characterized by inflammation. However, data regarding a polymorphism in the CRP gene, which has been reported in candidate gene analysis as an independent risk factor for stroke, have been conflicting [122, 123]. Other components of the inflammatory cascade studied include IL-1a, TNFα, IL-10, ICAM-1, though evidence has again been inconsistent [124–128] with studies limited by heterogeneity and small sample size.

Risk factors for AF: Hypertension, Obesity and Common Genetic Variants

Hypertension is an important modifiable risk factor for AF [129]. While GWAS for hypertension have not identified common SNPs that have also been associated with AF, several genetic variants associated with hypertension may contribute to AF. The genetic architecture of blood pressure appears to be extremely complex with ~100 independent variants accounting for only around 2% of the variance in blood pressure [130]. These include loci near *NPPA* and *NPPB*, which encode precursors of the natriuretic peptides. Polymorphisms in these loci correlate with the level of circulating natriuretic peptide and blood pressure [130, 131]. AF has been associated with variants in the *NPPA* gene [132, 133], causing resistance to ANP breakdown and increasing its half-life [37]. Atrial amyloid is largely derived from ANP. Transcriptomic differential gene expression analysis (discussed above) showed that AF susceptibility is associated with decreased targets of cellular stress

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

response to protein unfolding [9], leading to abnormal proteostasis and deposition of unfolded amyloid protein.

Several genes involved in the RAAS, including those for the angiotensin converting enzyme, angiotensin and angiotensin receptor, have been associated with essential hypertension by candidate gene analysis[134–136]. Some studies of patients with essential hypertension have identified RAAS gene variants that are related to increased risk of AF [73, 76, 137]. Polymorphisms in this system have also been linked to stroke. However, the RAAS system is involved in multiple pathologies and attributing some variants to certain disease states might be challenging in the setting of so many confounders. Nevertheless, the possibility that some variants contribute to multiple pathologies through pleiotropy seems plausible, as AF, obesity, hypertension are highly intertwined conditions and such pleiotropy has been well documented in other settings (Figure 1).

Other clinical risk factors for AF that have been associated with common genetic variants include obesity and the metabolic syndrome. Obesity increases the risk of atrial fibrillation by almost 50%[138]. In a GWAS meta-analysis of waist-hip ratio, *TBX15* was associated with increased waist-hip ratio. Unlike the other SNP hits in this study, *TBX15* did not differ by gender[139]. Candidate gene studies have shown that *CAV1* polymorphisms resulting in decreased expression of the gene are associated with insulin resistance and the metabolic syndrome in Hispanic and Caucasian cohorts[140, 141].

Conclusions

Genomic analyses have provided important insights into the pathogenesis of AF, though much still remains unknown. While GWAS have identified common genetic variants associated with AF, identification of the causative genetic variants and their functional mechanisms remains challenging. For example, the top replicated AF-associated locus on chromosome 4q25 is in an intergenic region, but the functional variants and the direct mechanisms by which these variants increase AF susceptibility is still poorly understood. The nearest gene, *PITX2*, is a biologically attractive candidate causative gene, as it directs asymmetric morphogenesis of the heart, is involved in the formation of pulmonary veins, and represses a left atrial sinus node program that might contribute to pulmonary vein triggers. However, the connections between the AF-associated SNPs and gene expression or function of *PITX2* remain elusive. Transcriptional studies in atrial tissue have made connections between SNPs and genes in several other AF risk loci, including for *MYOZ1, CAV1, C9orf3*, *CAND2, GJA1*, and *TBX5.* Analyses also suggest that the delay in AF manifestation until later in life may be associated with impaired transcriptional responses to cellular stress. Future GWAS of AF with much larger sample sizes will undoubtedly increase the number of susceptibility loci that are significantly associated with AF risk. Further studies are needed to determine functional mechanisms by which these genetic loci contribute to AF risk, stroke, and/or other common co-morbid conditions. Understanding these mechanisms will set the stage for the development of new therapeutic agents that target the causative genes.

Acknowledgments

This work was supported by the National Institutes of Health grant R01 HL111314 to MKC, the NIH National Center for Research Resources for Case Western Reserve University and Cleveland Clinic Clinical and Translational Science Award UL1-RR024989, the Cleveland Clinic Department of Cardiovascular Medicine philanthropy research funds, and the Tomsich Atrial Fibrillation Research Fund.

References

Papers of particular interest, published recently, have been highlighted as:

•Of importance

••Of major importance

1. Arnar DO, Thorvaldsson S, Manolio TA, et al. Eur Heart J. Mar.2006 27:708–712. [PubMed: 16428254]
2. Beckmann BM, Holinski-Feder E, Walter MC, et al. Laminopathy presenting as familial atrial fibrillation. Int J Cardiol. Nov 19.2010 145:394–396. [PubMed: 20472316]
3. Body SC, Collard CD, Shernan SK, et al. Variation in the 4q25 chromosomal locus predicts atrial fibrillation after coronary artery bypass graft surgery. Circ Cardiovasc Genet. Oct.2009 2:499–506. [PubMed: 20031626]
4. Benjamin EJ, Rice KM, Arking DE, et al. Variants in ZFHX3 are associated with atrial fibrillation in individuals of European ancestry. Nat Genet. Aug.2009 41:879–881. [PubMed: 19597492]
5. Brugada R, Tapscott T, Czernuszewicz GZ, et al. Identification of a genetic locus for familial atrial fibrillation. N Engl J Med. Mar 27.1997 336:905–911. [PubMed: 9070470]
6. Chen YH, Xu SJ, Bendahhou S, et al. KCNQ1 gain-of-function mutation in familial atrial fibrillation. Science. Jan 10.2003 299:251–254. [PubMed: 12522251]
7. Christophersen IE, Holmegard HN, Jabbari J, et al. Rare variants in GJA5 are associated with early-onset lone atrial fibrillation. Can J Cardiol. Jan.2013 29:111–116. [PubMed: 23040431]
8. Christophersen IE, Ellinor PT. Genetics of atrial fibrillation: from families to genomes. J Hum Genet. May 21.2015

9•. Deshmukh A, Barnard J, Sun H, et al. Left atrial transcriptional changes associated with atrial fibrillation susceptibility and persistence. Circ Arrhythm Electrophysiol. Feb.2015 8:32–41. Differential gene expression by atrial rhythm status showed AF was associated with decreased transcriptional responses to cellular stress. [PubMed: 25523945]

10•. Ellinor PT, Lunetta KL, Albert CM, et al. Meta-analysis identifies six new susceptibility loci for atrial fibrillation. Nat Genet. Apr 29.2012 44:670–675. Large GWAS meta-analysis identified new AF risk loci. [PubMed: 22544366]

11. Ellinor PT, Lunetta KL, Glazer NL, et al. Common variants in KCNN3 are associated with lone atrial fibrillation. Nat Genet. Mar.2010 42:240–244. [PubMed: 20173747]
12. Gudbjartsson DF, Arnar DO, Helgadottir A, et al. Variants conferring risk of atrial fibrillation on chromosome 4q25. Nature. Jul 19.2007 448:353–357. [PubMed: 17603472]
13. Hayashi K, Konno T, Tada H, et al. Functional Characterization of Rare Variants Implicated in Susceptibility to Lone Atrial Fibrillation. Circ Arrhythm Electrophysiol. Jun 30.2015
14. Lee KT, Yeh HY, Tung CP, et al. Association of RS2200733 but not RS10033464 on 4q25 with atrial fibrillation based on the recessive model in a Taiwanese population. Cardiology. 2010; 116:151–156. [PubMed: 20606429]

15•. Sinner MF, Tucker NR, Lunetta KL, et al. Integrating genetic, transcriptional, and functional analyses to identify 5 novel genes for atrial fibrillation. Circulation. Oct 7.2014 130:1225–1235. Combines transcriptional analysis and functional analysis to give some information regarding the possible mechanism by which a risk SNP could lead to AF. [PubMed: 25124494]

16. Thibodeau IL, Xu J, Li Q, et al. Paradigm of genetic mosaicism and lone atrial fibrillation: physiological characterization of a connexin 43-deletion mutant identified from atrial tissue. Circulation. Jul 20.2010 122:236–244. [PubMed: 20606116]

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

17. Lloyd-Jones DM, Wang TJ, Leip EP, et al. Lifetime risk for development of atrial fibrillation: the Framingham Heart Study. Circulation. Aug 31.2004 110:1042–1046. [PubMed: 15313941]
18. Schnabel RB, Yin X, Gona P, et al. 50 year trends in atrial fibrillation prevalence, incidence, risk factors, and mortality in the Framingham Heart Study: a cohort study. Lancet. Jul 11.2015 386:154–162. [PubMed: 25960110]
19. Fox CS, Parise H, D'Agostino S, et al. Parental atrial fibrillation as a risk factor for atrial fibrillation in offspring. JAMA. Jun 16.2004 291:2851–2855. [PubMed: 15199036]
20. Lubitz SA, Yin X, Fontes JD, et al. Association between familial atrial fibrillation and risk of new-onset atrial fibrillation. JAMA. Nov 24.2010 304:2263–2269. [PubMed: 21076174]
21. Ellinor PT, Yoerger DM, Ruskin JN, et al. Familial aggregation in lone atrial fibrillation. Hum Genet. Nov.2005 118:179–184. [PubMed: 16133178]
22. Christophersen IE, Ravn LS, Budtz-Joergensen E, et al. Familial aggregation of atrial fibrillation: a study in Danish twins. Circ Arrhythm Electrophysiol. Aug.2009 2:378–383. [PubMed: 19808493]
23. Bertram H, Paul T, Beyer F, et al. Familial idiopathic atrial fibrillation with bradyarrhythmia. Eur J Pediatr. Jan.1996 155:7–10. [PubMed: 8750801]
24. Brugada R. Atrial fibrillation: is the solution in the genes? Rev Esp Cardiol. Jun.1997 50:367. [PubMed: 9304158]
25. Levy R. Paroxysmal auricular fibrillation and flutter without signs of organic cardiac disease in two brothers. J Mt Sinai Hosp. 1942; 8:765–770.
26. Olson TM, Alekseev AE, Liu XK, et al. Kv1.5 channelopathy due to KCNA5 loss-of-function mutation causes human atrial fibrillation. Hum Mol Genet. Jul 15.2006 15:2185–2191. [PubMed: 16772329]
27. Li RG, Wang Q, Xu YJ, et al. Mutations of the SCN4B-encoded sodium channel beta4 subunit in familial atrial fibrillation. Int J Mol Med. Jul.2013 32:144–150. [PubMed: 23604097]
28. Otway R, Vandenberg JI, Guo G, et al. Stretch-sensitive KCNQ1 mutation A link between genetic and environmental factors in the pathogenesis of atrial fibrillation? J Am Coll Cardiol. Feb 6.2007 49:578–586. [PubMed: 17276182]
29. Das S, Makino S, Melman YF, Shea MA, et al. Mutation in the S3 segment of KCNQ1 results in familial lone atrial fibrillation. Heart Rhythm. Aug.2009 6:1146–1153. [PubMed: 19632626]
30. Yang Y, Xia M, Jin Q, et al. Identification of a KCNE2 gain-of-function mutation in patients with familial atrial fibrillation. Am J Hum Genet. Nov.2004 75:899–905. [PubMed: 15368194]
31. Hong K, Bjerregaard P, Gussak I, et al. Short QT syndrome and atrial fibrillation caused by mutation in KCNH2. J Cardiovasc Electrophysiol. Apr.2005 16:394–396. [PubMed: 15828882]
32. Xia M, Jin Q, Bendahhou S, et al. A Kir2.1 gain-of-function mutation underlies familial atrial fibrillation. Biochem Biophys Res Commun. Jul 15.2005 332:1012–1019. [PubMed: 15922306]
33. Makiyama T, Akao M, Shizuta S, et al. A novel SCN5A gain-of-function mutation M1875T associated with familial atrial fibrillation. J Am Coll Cardiol. Oct 14.2008 52:1326–1334. [PubMed: 18929244]
34. Benito B, Brugada R, Perich RM, et al. A mutation in the sodium channel is responsible for the association of long QT syndrome and familial atrial fibrillation. Heart Rhythm. Oct.2008 5:1434–1440. [PubMed: 18929331]
35. Hong K, Piper DR, Diaz-Valdecantos A, et al. De novo KCNQ1 mutation responsible for atrial fibrillation and short QT syndrome in utero. Cardiovasc Res. Dec 1.2005 68:433–440. [PubMed: 16109388]
36. Gollob MH, Green MS, Tang AS, et al. Identification of a gene responsible for familial Wolff-Parkinson-White syndrome. N Engl J Med. Jun 14.2001 344:1823–1831. [PubMed: 11407343]
37. Hodgson-Zingman DM, Karst ML, Zingman LV, et al. Atrial natriuretic peptide frameshift mutation in familial atrial fibrillation. N Engl J Med. Jul 10.2008 359:158–165. [PubMed: 18614783]
38. Wang J, Sun YM, Yang YQ. Mutation spectrum of the GATA4 gene in patients with idiopathic atrial fibrillation. Mol Biol Rep. Aug.2012 39:8127–8135. [PubMed: 22552926]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

39. Posch MG, Boldt LH, Polotzki M, et al. Mutations in the cardiac transcription factor GATA4 in patients with lone atrial fibrillation. Eur J Med Genet. Jul-Aug;2010 53:201–203. [PubMed: 20363377]

40. Wang XH, Huang CX, Wang Q, et al. A novel GATA5 loss-of-function mutation underlies lone atrial fibrillation. Int J Mol Med. Jan.2013 31:43–50. [PubMed: 23175127]

41. Huang RT, Xue S, Xu YJ, et al. A novel NKX2.5 loss-of-function mutation responsible for familial atrial fibrillation. Int J Mol Med. May.2013 31:1119–1126. [PubMed: 23525379]

42. Gollob MH, Jones DL, Krahn AD, et al. Somatic mutations in the connexin 40 gene (GJA5) in atrial fibrillation. N Engl J Med. Jun 22.2006 354:2677–2688. [PubMed: 16790700]

43. Di Pino A, Caruso E, Costanzo L, et al. A novel RyR2 mutation in a 2-year-old baby presenting with atrial fibrillation, atrial flutter, and atrial ectopic tachycardia. Heart Rhythm. Aug.2014 11:1480–1483. [PubMed: 24793461]

44. Kazemian P, Gollob MH, Pantano A, et al. A novel mutation in the RYR2 gene leading to catecholaminergic polymorphic ventricular tachycardia and paroxysmal atrial fibrillation: dose-dependent arrhythmia-event suppression by beta-blocker therapy. Can J Cardiol. Nov-Dec;2011 27:870.e7–870.10.

45. Pan H, Richards AA, Zhu X, et al. A novel mutation in LAMIN A/C is associated with isolated early-onset atrial fibrillation and progressive atrioventricular block followed by cardiomyopathy and sudden cardiac death. Heart Rhythm. May.2009 6:707–710. [PubMed: 19328042]

46. Oberti C, Wang L, Li L, et al. Genome-wide linkage scan identifies a novel genetic locus on chromosome 5p13 for neonatal atrial fibrillation associated with sudden death and variable cardiomyopathy. Circulation. Dec 21.2004 110:3753–3759. [PubMed: 15596564]

47. Zhang X, Chen S, Yoo S, et al. Mutation in nuclear pore component NUP155 leads to atrial fibrillation and early sudden cardiac death. Cell. Dec 12.2008 135:1017–1027. [PubMed: 19070573]

48. Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. Nature. Oct 8.2009 461:747–753. [PubMed: 19812666]

49. Ioannidis JP, Ntzani EE, Trikalinos TA, et al. Replication validity of genetic association studies. Nat Genet. Nov.2001 29:306–309. [PubMed: 11600885]

50. Tabor HK, Risch NJ, Myers RM. Candidate-gene approaches for studying complex genetic traits: practical considerations. Nat Rev Genet. May.2002 3:391–397. [PubMed: 11988764]

51. Li L, Shen C, Yao Z, et al. Genetic variants of potassium voltage-gated channel genes (KCNQ1, KCNH2, and KCNE1) affected the risk of atrial fibrillation in elderly patients. Genet Test Mol Biomarkers. Jul.2015 19:359–365. [PubMed: 26066992]

52. Ravn LS, Aizawa Y, Pollevick GD, et al. Gain of function in IKs secondary to a mutation in KCNE5 associated with atrial fibrillation. Heart Rhythm. Mar.2008 5:427–435. [PubMed: 18313602]

53. Mao T, Miao HJ, Xu GJ, et al. Association of single nucleotide polymorphism of KCNE1 and KCNE4 gene with atrial fibrillation in Xinjiang Uygur and Han population. Zhonghua Xin Xue Guan Bing Za Zhi. Nov.2013 41:916–921. [PubMed: 24370217]

54. Lundby A, Ravn LS, Svendsen JH, et al. KCNE3 mutation V17M identified in a patient with lone atrial fibrillation. Cell Physiol Biochem. 2008; 21:47–54. [PubMed: 18209471]

55. Jabbari J, Olesen MS, Holst AG, et al. Common polymorphisms in KCNJ5 [corrected] are associated with early-onset lone atrial fibrillation in Caucasians. Cardiology. 2011; 118:116–120. [PubMed: 21555883]

56. Delaney JT, Muhammad R, Blair MA, et al. A KCNJ8 mutation associated with early repolarization and atrial fibrillation. Europace. Oct.2012 14:1428–1432. [PubMed: 22562657]

57. Olesen MS, Refsgaard L, Holst AG, et al. A novel KCND3 gain-of-function mutation associated with early-onset of persistent lone atrial fibrillation. Cardiovasc Res. Jun 1.2013 98:488–495. [PubMed: 23400760]

58. Nielsen JB, Bentzen BH, Olesen MS, et al. Gain-of-function mutations in potassium channel subunit KCNE2 associated with early-onset lone atrial fibrillation. Biomark Med. 2014; 8:557–570. [PubMed: 24796621]

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

59. Ma KJ, Li N, Teng SY, et al. Modulation of KCNQ1 current by atrial fibrillation-associated KCNE4 (145E/D) gene polymorphism. Chin Med J (Engl). Jan 20.2007 120:150–154. [PubMed: 17335661]
60. Lemoine MD, Duverger JE, Naud P, et al. Arrhythmogenic left atrial cellular electrophysiology in a murine genetic long QT syndrome model. Cardiovasc Res. Oct 1.2011 92:67–74. [PubMed: 21672931]
61. Christophersen IE, Olesen MS, Liang B, et al. Genetic variation in KCNA5: impact on the atrial-specific potassium current IKur in patients with lone atrial fibrillation. Eur Heart J. May.2013 34:1517–1525. [PubMed: 23264583]
62. Darbar D, Kannankeril PJ, Donahue BS, et al. Cardiac sodium channel (SCN5A) variants associated with atrial fibrillation. Circulation. Apr 15.2008 117:1927–1935. [PubMed: 18378609]
63. Li Q, Huang H, Liu G, Lam K, et al. Gain-of-function mutation of Nav1.5 in atrial fibrillation enhances cellular excitability and lowers the threshold for action potential firing. Biochem Biophys Res Commun. Feb 27.2009 380:132–137. [PubMed: 19167345]
64. Watanabe H, Darbar D, Kaiser DW, et al. Mutations in sodium channel beta1- and beta2-subunits associated with atrial fibrillation. Circ Arrhythm Electrophysiol. Jun.2009 2:268–275. [PubMed: 19808477]
65. Olesen MS, Jespersen T, Nielsen JB, et al. Mutations in sodium channel beta-subunit SCN3B are associated with early-onset lone atrial fibrillation. Cardiovasc Res. Mar 1.2011 89:786–793. [PubMed: 21051419]
66. Savio-Galimberti E, Weeke P, Muhammad R, et al. SCN10A/Nav1.8 modulation of peak and late sodium currents in patients with early onset atrial fibrillation. Cardiovasc Res. Nov 1.2014 104:355–363. [PubMed: 25053638]
67. Wirka RC, Gore S, Van Wagoner DR, et al. A common connexin-40 gene promoter variant affects connexin-40 expression in human atria and is associated with atrial fibrillation. Circ Arrhythm Electrophysiol. Feb.2011 4:87–93. [PubMed: 21076161]
68. Juang JM, Chern YR, Tsai CT, et al. The association of human connexin 40 genetic polymorphisms with atrial fibrillation. Int J Cardiol. Mar 2.2007 116:107–112. [PubMed: 16814413]
69. Sun Y, Tong X, Chen H, et al. An atrial-fibrillation-linked connexin40 mutant is retained in the endoplasmic reticulum and impairs the function of atrial gap-junction channels. Dis Model Mech. May.2014 7:561–569. [PubMed: 24626989]
70. Xie WH, Chang C, Xu YJ, et al. Prevalence and spectrum of Nkx2.5 mutations associated with idiopathic atrial fibrillation. Clinics (Sao Paulo). Jun.2013 68:777–784. [PubMed: 23778487]
71. Yu H, Xu JH, Song HM, et al. Mutational spectrum of the NKX2-5 gene in patients with lone atrial fibrillation. Int J Med Sci. Apr 7.2014 11:554–563. [PubMed: 24782644]
72. Brauch KM, Chen LY, Olson TM. Comprehensive mutation scanning of LMNA in 268 patients with lone atrial fibrillation. Am J Cardiol. May 15.2009 103:1426–1428. [PubMed: 19427440]
73. Jiang MH, Su YM, Tang JZ, et al. Angiotensin-converting enzyme gene 2350 G/A polymorphism and susceptibility to atrial fibrillation in Han Chinese patients with essential hypertension. Clinics (Sao Paulo). Nov.2013 68:1428–1432. [PubMed: 24270955]
74. Ravn LS, Benn M, Nordestgaard BG, et al. Angiotensinogen and ACE gene polymorphisms and risk of atrial fibrillation in the general population. Pharmacogenet Genomics. Jun.2008 18:525–533. [PubMed: 18496132]
75. Zhao LQ, Wen ZJ, Wei Y, et al. Polymorphisms of renin-angiotensin-aldosterone system gene in chinese han patients with nonfamilial atrial fibrillation. PLoS One. Feb 27.2015 10:e0117489. [PubMed: 25723521]
76. Tsai CT, Lai LP, Lin JL, et al. Renin-angiotensin system gene polymorphisms and atrial fibrillation. Circulation. Apr 6.2004 109:1640–1646. [PubMed: 15023884]
77. Ehret GB. Genome-wide association studies: contribution of genomics to understanding blood pressure and essential hypertension. Curr Hypertens Rep. Feb.2010 12:17–25. [PubMed: 20425154]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

78. Husser D, Adams V, Piorkowski C, et al. Chromosome 4q25 variants and atrial fibrillation recurrence after catheter ablation. J Am Coll Cardiol. Feb 23.2010 55:747–753. [PubMed: 20170812]
79. Lubitz SA, Ozcan C, Magnani JW, et al. Genetics of atrial fibrillation: implications for future research directions and personalized medicine. Circ Arrhythm Electrophysiol. Jun.2010 3:291–299. [PubMed: 20551423]
80. Liu X, Wang F, Knight AC, Zhao J, et al. Common variants for atrial fibrillation: results from genome-wide association studies. Hum Genet. Jan.2012 131:33–39. [PubMed: 21735173]
81. Lubitz SA, Sinner MF, Lunetta KL, et al. Independent susceptibility markers for atrial fibrillation on chromosome 4q25. Circulation. Sep 7.2010 122:976–984. [PubMed: 20733104]
82. Liu C, Liu W, Lu MF, et al. Regulation of left-right asymmetry by thresholds of Pitx2c activity. Development. Jun.2001 128:2039–2048. [PubMed: 11493526]
83. Wang J, Klysik E, Sood S, et al. Pitx2 prevents susceptibility to atrial arrhythmias by inhibiting left-sided pacemaker specification. Proc Natl Acad Sci U S A. May 25.2010 107:9753–9758. [PubMed: 20457925]
84. Mommersteeg MT, Brown NA, Prall OW, et al. Pitx2c and Nkx2-5 are required for the formation and identity of the pulmonary myocardium. Circ Res. Oct 26.2007 101:902–909. [PubMed: 17823370]
85. Kirchhof P, Kahr PC, Kaese S, et al. PITX2c is expressed in the adult left atrium, and reducing Pitx2c expression promotes atrial fibrillation inducibility and complex changes in gene expression. Circ Cardiovasc Genet. Apr.2011 4:123–133. [PubMed: 21282332]
86. Franco D, Chinchilla A, Daimi H, et al. Modulation of conductive elements by Pitx2 and their impact on atrial arrhythmogenesis. Cardiovasc Res. Jul 15.2011 91:223–231. [PubMed: 21427120]
87. Chinchilla A, Daimi H, Lozano-Velasco E, et al. PITX2 insufficiency leads to atrial electrical and structural remodeling linked to arrhythmogenesis. Circ Cardiovasc Genet. Jun.2011 4:269–279. [PubMed: 21511879]
88. Lozano-Velasco E, Hernandez-Torres F, Daimi H, et al. Pitx2 impairs calcium handling in a dose-dependent manner by modulating Wnt signalling. Cardiovasc Res. Jan 1.2016 109:55–66. [PubMed: 26243430]
89. Wang J, Bai Y, Li N, et al. Pitx2-microRNA pathway that delimits sinoatrial node development and inhibits predisposition to atrial fibrillation. Proc Natl Acad Sci U S A. Jun 24.2014 111:9181–9186. [PubMed: 24927531]
90. Haissaguerre M, Jais P, Shah DC, et al. Spontaneous initiation of atrial fibrillation by ectopic beats originating in the pulmonary veins. N Engl J Med. Sep 3.1998 339:659–666. [PubMed: 9725923]
91. Tao Y, Zhang M, Li L, et al. Pitx2, an atrial fibrillation predisposition gene, directly regulates ion transport and intercalated disc genes. Circ Cardiovasc Genet. Feb.2014 7:23–32. [PubMed: 24395921]
92. Ihida-Stansbury K, McKean DM, Gebb SA, et al. Paired-related homeobox gene Prx1 is required for pulmonary vascular development. Circ Res. Jun 11.2004 94:1507–1514. [PubMed: 15117820]
93. Skibsbye L, Poulet C, Diness JG, et al. Small-conductance calcium-activated potassium (SK) channels contribute to action potential repolarization in human atria. Cardiovasc Res. Jul 1.2014 103:156–167. [PubMed: 24817686]
94. Macri V, Mahida SN, Zhang ML, et al. A novel trafficking-defective HCN4 mutation is associated with early-onset atrial fibrillation. Heart Rhythm. Jun.2014 11:1055–1062. [PubMed: 24607718]
95. Martin RI, Babaei MS, Choy MK, et al. Genetic variants associated with risk of atrial fibrillation regulate expression of PITX2, CAV1, MYOZ1, C9orf3 and FANCC. J Mol Cell Cardiol. Aug.2015 85:207–214. [PubMed: 26073630]
96. Yi SL, Liu XJ, Zhong JQ, et al. Role of caveolin-1 in atrial fibrillation as an anti-fibrotic signaling molecule in human atrial fibroblasts. PLoS One. Jan 14.2014 9:e85144. [PubMed: 24454806]
97. Albert FW, Kruglyak L. The role of regulatory variation in complex traits and disease. Nat Rev Genet. Apr.2015 16:197–212. [PubMed: 25707927]
98. Lin H, Dolmatova EV, Morley MP, et al. Gene expression and genetic variation in human atria. Heart Rhythm. Feb.2014 11:266–271. [PubMed: 24177373]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

99. Gore-Panter SR, Hsu J, Hanna P, et al. Atrial Fibrillation associated chromosome 4q25 variants are not associated with PITX2c expression in human adult left atrial appendages. PLoS One. Jan 22.2014 9:e86245. [PubMed: 24465984]

100. Gore-Panter SR, Hsu J, Barnard J, et al. PANCR, the PITX2 Adjacent Noncoding RNA, Is Expressed in Human Left Atria and Regulates PITX2c Expression. Circ Arrhythm Electrophysiol. 2016; 9

101. Suzuki H, Suzuki A, Maekawa Y, et al. Interplay between two myogenesis-related proteins: TBP-interacting protein 120B and MyoD. Gene. Aug 10.2012 504:213–219. [PubMed: 22613845]

102. Gemel J, Levy AE, Simon AR, et al. Connexin40 abnormalities and atrial fibrillation in the human heart. J Mol Cell Cardiol. Nov.2014 76:159–168. [PubMed: 25200600]

103. Pal R, Ramdzan ZM, Kaur S, et al. CUX2 Protein Functions as an Accessory Factor in the Repair of Oxidative DNA Damage. J Biol Chem. Sep 11.2015 290:22520–22531. [PubMed: 26221032]

104. Rocken C, Peters B, Juenemann G, et al. Atrial amyloidosis: an arrhythmogenic substrate for persistent atrial fibrillation. Circulation. Oct 15.2002 106:2091–2097. [PubMed: 12379579]

105. Leone O, Boriani G, Chiappini B, et al. Amyloid deposition as a cause of atrial remodelling in persistent valvular atrial fibrillation. Eur Heart J. Jul.2004 25:1237–1241. [PubMed: 15246642]

106. Linke RP, Voigt C, Storkel FS, et al. N-terminal amino acid sequence analysis indicates that isolated atrial amyloid is derived from atrial natriuretic peptide. Virchows Arch B Cell Pathol Incl Mol Pathol. 1988; 55:125–127. [PubMed: 2901160]

107. Ma, J-F., Yang, F., Mahida, et al. TBX5 mutations contribute to early-onset atrial fibrillation in Chinese and Caucasians. Cardiovascular Research. 2016. pp. http://cardiovascres.oxfordjournals.org/content/early/2016/01/12/cvr.cvw003.abstract

108. Souto JC, Pena G, Ziyatdinov A, et al. A genomewide study of body mass index and its genetic correlation with thromboembolic risk. Results from the GAIT project. Thromb Haemost. Nov. 2014 112:1036–1043. [PubMed: 25118907]

109. Adams HP Jr, Bendixen BH, Kappelle LJ, et al. Classification of subtype of acute ischemic stroke. Definitions for use in a multicenter clinical trial. TOAST. Trial of Org 10172 in Acute Stroke Treatment. Stroke. Jan.1993 24:35–41. [PubMed: 7678184]

110. Gretarsdottir S, Thorleifsson G, Manolescu A, et al. Risk variants for atrial fibrillation on chromosome 4q25 associate with ischemic stroke. Ann Neurol. Oct.2008 64:402–409. [PubMed: 18991354]

111. Wnuk M, Pera J, Jagiella J, et al. The rs2200733 variant on chromosome 4q25 is a risk factor for cardioembolic stroke related to atrial fibrillation in Polish patients. Neurol Neurochir Pol. Mar-Apr;2011 45:148–152. [PubMed: 21574119]

112. Lemmens R, Buysschaert I, Geelen V, et al. The association of the 4q25 susceptibility variant for atrial fibrillation with stroke is limited to stroke of cardioembolic etiology. Stroke. Sep.2010 41:1850–1857. [PubMed: 20671249]

113. Gudbjartsson DF, Holm H, Gretarsdottir S, et al. A sequence variant in ZFHX3 on 16q22 associates with atrial fibrillation and ischemic stroke. Nat Genet. Aug.2009 41:876–878. [PubMed: 19597491]

114. Traylor M, Farrall M, Holliday EG, et al. Genetic risk factors for ischaemic stroke and its subtypes (the METASTROKE collaboration): a meta-analysis of genome-wide association studies. Lancet Neurol. Nov.2012 11:951–962. [PubMed: 23041239]

115. Auer PL, Nalls M, Meschia JF, et al. Rare and Coding Region Genetic Variants Associated With Risk of Ischemic Stroke: The NHLBI Exome Sequence Project. JAMA Neurol. Jul.2015 72:781–788. [PubMed: 25961151]

116•. Tada H, Shiffman D, Smith JG, et al. Twelve-single nucleotide polymorphism genetic risk score identifies individuals at increased risk for future atrial fibrillation and stroke. Stroke. Oct.2014 45:2856–2862. Genetic risk score composed of AF risk SNPs identified subjects at higher risk for future AF and stroke. [PubMed: 25123217]

117. Schmidt H, Aulchenko YS, Schweighofer N, et al. Angiotensinogen promoter B-haplotype associated with cerebral small vessel disease enhances basal transcriptional activity. Stroke. Nov. 2004 35:2592–2597. [PubMed: 15472104]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

118. Saidi S, Mallat SG, Almawi WY, et al. Association between renin-angiotensin-aldosterone system genotypes and haplotypes and risk of ischemic stroke of atherosclerotic etiology. Acta Neurol Scand. Jun.2009 119:356–363. [PubMed: 19473121]
119. Tsai CT, Chang SN, Chang SH, et al. Renin-angiotensin system gene polymorphisms predict the risk of stroke in patients with atrial fibrillation: a 10-year prospective follow-up study. Heart Rhythm. Aug.2014 11:1384–1390. [PubMed: 24732371]
120. Zhao J, Qin X, Li S, et al. Association between the ACE I/D polymorphism and risk of ischemic stroke: an updated meta-analysis of 47,026 subjects from 105 case-control studies. J Neurol Sci. Oct 15.2014 345:37–47. [PubMed: 25082780]
121. Stetskaia TA, Bushueva OI, Bulgakova IV, et al. Association of T174M polymorphism of the angiotensinogen gene with the higher risk of cerebral stroke in women. Ter Arkh. 2014; 86:66–71.
122. Liu Y, Geng PL, Yan FQ, et al. C-reactive Protein −717A>G and −286C>T>A Gene Polymorphism and Ischemic Stroke. Chin Med J (Engl). Jun 20.2015 128:1666–1670. [PubMed: 26063371]
123. Wang Q, Ding H, Tang JR, et al. C-reactive protein polymorphisms and genetic susceptibility to ischemic stroke and hemorrhagic stroke in the Chinese Han population. Acta Pharmacol Sin. Mar.2009 30:291–298. [PubMed: 19262552]
124. Niu YM, Weng H, Zhang C, et al. Systematic Review by Multivariate Meta-analyses on the Possible Role of Tumor Necrosis Factor-alpha Gene Polymorphisms in Association with Ischemic Stroke. Neuromolecular Med. Aug 1.2015
125. Zhao N, Liu X, Wang Y, et al. Association of inflammatory gene polymorphisms with ischemic stroke in a Chinese Han population. J Neuroinflammation. Jul 6.2012 9:162-2094-9-162. [PubMed: 22769019]
126. Jin J, Li W, Peng L, et al. Relationship between interleukin-10 −1082A/G polymorphism and risk of ischemic stroke: a meta-analysis. PLoS One. Apr 14.2014 9:e94631. [PubMed: 24732312]
127. Muino E, Krupinski J, Carrera C, et al. An Inflammatory Polymorphisms Risk Scoring System for the Differentiation of Ischemic Stroke Subtypes. Mediators Inflamm. 2015; 2015:569714. [PubMed: 26355258]
128. Li J, Song J, Jiang MH, et al. Interleukin-6 promoter polymorphisms and susceptibility to atrial fibrillation in elderly Han Chinese patients with essential hypertension. J Interferon Cytokine Res. Nov.2012 32:542–547. [PubMed: 22924939]
129. Huxley RR, Lopez FL, Folsom AR, et al. Absolute and attributable risks of atrial fibrillation in relation to optimal and borderline risk factors: the Atherosclerosis Risk in Communities (ARIC) study. Circulation. Apr 12.2011 123:1501–1508. [PubMed: 21444879]
130. International Consortium for Blood Pressure Genome-Wide Association Studies. Ehret GB, Munroe PB, et al. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. Nature. Sep 11.2011 478:103–109. [PubMed: 21909115]
131. Newton-Cheh C, Larson MG, Vasan RS, et al. Association of common variants in NPPA and NPPB with circulating natriuretic peptides and blood pressure. Nat Genet. Mar.2009 41:348–353. [PubMed: 19219041]
132. Roberts JD, Davies RW, Lubitz SA, et al. Evaluation of non-synonymous NPPA single nucleotide polymorphisms in atrial fibrillation. Europace. Aug.2010 12:1078–1083. [PubMed: 20543198]
133. Ren X, Xu C, Zhan C, et al. Identification of NPPA variants associated with atrial fibrillation in a Chinese GeneID population. Clin Chim Acta. Apr 2.2010 411:481–485. [PubMed: 20064500]
134. Yang YL, Mo YP, He YS, et al. Correlation between renin-angiotensin system gene polymorphisms and essential hypertension in the Chinese Yi ethnic group. J Renin Angiotensin Aldosterone Syst. Aug 17.2015
135. Cosenso-Martin LN, Vaz-de-Melo RO, Pereira LR, et al. Angiotensin-converting enzyme insertion/deletion polymorphism, 24-h blood pressure profile and left ventricular hypertrophy in hypertensive individuals: a cross-sectional study. Eur J Med Res. Sep 4.2015 20:74-015-0166-9. [PubMed: 26336879]

136. Li Q, Sun L, Du J, et al. Risk given by AGT polymorphisms in inducing susceptibility to essential hypertension among isolated populations from a remote region of China: A case-control study among the isolated populations. J Renin Angiotensin Aldosterone Syst. Sep 20.2015

137. Sun X, Yang J, Hou X, et al. Relationship between −344T/C polymorphism in the aldosterone synthase gene and atrial fibrillation in patients with essential hypertension. J Renin Angiotensin Aldosterone Syst. Dec.2011 12:557–563. [PubMed: 21846681]

138. Wanahita N, Messerli FH, Bangalore S, et al. Atrial fibrillation and obesity–results of a meta-analysis. Am Heart J. Feb.2008 155:310–315. [PubMed: 18215602]

139. Heid IM, Jackson AU, Randall JC, et al. Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution. Nat Genet. Nov.2010 42:949–960. [PubMed: 20935629]

140. Baudrand R, Goodarzi MO, Vaidya A, et al. A prevalent caveolin-1 gene variant is associated with the metabolic syndrome in Caucasians and Hispanics. Metabolism. Dec.2015 64:1674–1681. [PubMed: 26475177]

141. Pojoga LH, Underwood PC, Goodarzi MO, et al. Variants of the caveolin-1 gene: a translational investigation linking insulin resistance and hypertension. J Clin Endocrinol Metab. Aug.2011 96:E1288–92. [PubMed: 21613355]

142. Olson TM, Alekseev AE, Moreau C, et al. KATP channel mutation confers risk for vein of Marshall adrenergic atrial fibrillation. Nat Clin Pract Cardiovasc Med. Feb.2007 4:110–116. [PubMed: 17245405]

143. Jabbari J, Olesen MS, Yuan L, et al. Common and rare variants in SCN10A modulate the risk of atrial fibrillation. Circ Cardiovasc Genet. Feb.2015 8:64–73. [PubMed: 25691686]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

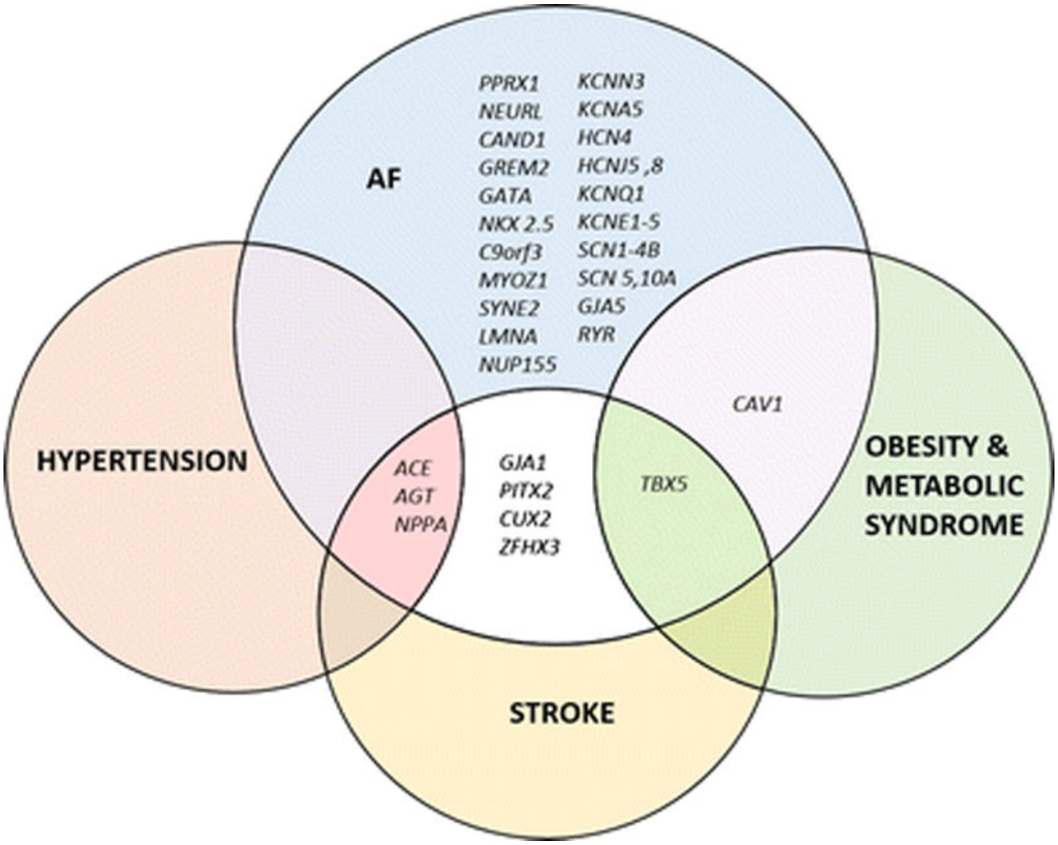

AF
PPRX1
NEURL
CAND1
GREM2
GATA
NKX 2.5
C9orf3
MYOZ1
SYNE2
LMNA
NUP155
KCNN3
KCNA5
HCN4
HCNJ5 ,8
KCNQ1
KCNE1-5
SCN1-4B
SCN 5,10A
GJA5
RYR
HYPERTENSION
ACE
AGT
NPPA
GJA1
PITX2
CUX2
ZFHX3
CAV1
TBX5
OBESITY & METABOLIC SYNDROME
STROKE

Figure 1.

Common implicated genes associated with AF, stroke, hypertension, and obesity/metabolic syndrome.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 1

Gene loci associated with atrial fibrillation

Gene	Function	Reference
Familial studies		
KCNQ1	Gain of function of potassium channel contributing to I_{ks}	[6, 28, 29, 35]
KCNE2	Gain of function mutation of the potassium channel responsible for the I_{Ks} current	[30]
KCNH2	Encodes for the channel responsible for the rapidly depolarizing current I_{kr}	[31]
KCNJ2	Encodes for the inward rectifier potassium channel Kir 2.1	[32]
KCNA5	Modulation of ultrarapid depolarizing current I_{Kur}	[26]
SCN5A	Nav 1.5 responsible for upstroke of action potential	[33, 34]
SCN4B	β subunit of the voltage gated sodium channel	[27]
PRKAG2	ϒ2 subunit of AMP-activated protein kinase which regulates ATP generation and use.	[36]
NPPA	Frameshift mutation causes ANP to be resistant to breakdown increasing its half life	[37]
ABCC9	ATP binding cassette leads to loss of function of I_{KATP}	[142]
RYR	Alteration of ryanodine receptor leading to imbalance of calcium homeostasis	[43, 44]
NUP155	Nucleoporin 155, a component of the nucleopore formation reducing nuclear envelope permeability	[47]
LMNA	Laminin A/C in inner nuclear membrane	[2, 45]
GATA 4,5	Zinc finger transcription factor involved in cardiac development	[38–40]
Candidate Gene studies		
KCNE 1,3,4 and *5*	Mutation of the βsubunit of voltage gated potassium channel leading to altered function of I_{Ks}	[51–54]
KCNJ 5 and *8*	α subunits of inwardly rectifying potassium channels	[55, 56]
KCND3	K_v4.3 α subunit causing increase in the transient outward potassium current I_{to}	[57]
SCN1B,2B,3B	Mutation in the β subunits of the sodium channel leading to decreased sodium current	[13, 64, 65]
SCN10A	Nav 1.8 which participates in the late sodium current	[66, 143]
GJA5	Connexin 40 in gap junctions altering action potential propagation	[7, 67, 68]
NKX2.5	Homeobox transcription factor involved in cardiogenesis	[70, 71]
RAAS	Angiotensin conversion enzyme inhibitor, angiotensin gene promoter and angiotensinogen polymorphisms	[74–76]
Genome Wide Association Studies		
PITX2	Right-left asymmetry, atrial cardiomyocyte and SA node development, intercalated disk?	[12]
ZFHX3	zinc finger homeobox 3; regulation of growth and differentiation of skeletal muscle and neuronal tissue.	[4, 113]
PRRX1	a homeodomain transcription factor expressed in connective tissue in the developing heart	[10]
KCNN3	Intronic variant in the gene encoding for the calcium activated potassium channel SK3	[11]
HCN4	hyperpolarization-activated cyclic nucleotide- gated 4 channel of the I_f	[10]
CAV1	Caveolin-1, a structural component of caveolae	
SYNE2	Intronic SNP encoding for nespirin 2; anchors the nucleus to the cytoskeleton	[10]
MYOZ1/SYNPO2L	Intergenic variant. *MYOZ1* encodes for myozenin 1 a protein involved in stabilizing the sarcomere	[10]
C9orf3	Opening reading frame of chromosome 9, potential causal gene is still unclear	[10]
GJA1	Connexin 43 in gap junctions altering action potential propagation	[15]
NEURL	E3 ubiquitin ligase- interacts with *PITX2* and leads to increased AP duration	[15]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Gene	Function	Reference
CAND2	TBP- interacting protein involved in myogenesis	[15]
TBX5	Transcription factor involved in the development of cardiac conduction system	[15]
CUX2	Cut-like homebox 2 a transcription factor involved in neural development	[15]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 2

Common variants between AF and Stroke

Reference	Number of patients	Nearby gene-SNP variant	Association
Lemmens et al[112]	4199 patients with IS	*PITX2*-rs1906591	AF and IS of CE etiology
Wnuk et al[111]	729 Polish patients	*PITX2*-rs2200733	CE stroke related to AF
Gretarsdottir et al[110]***	1661 Icelandic patients	*PITX2*-rs2200733	CE stroke only
Taylor et al [114] **	Meta-analysis	*PITX2*-rs6843082	CES
Gudbjartsson et al [113]***	including 74,393 patients	*ZFHX3*-rs7193343	IS and CES
Sinner et al. [15] **	Meta-analysis and metastroke 74,393 subjects	*CUX2*-rs6490029 *GJA1*-rs13216675 *TBX5*-rs10507248	IS CES CES
Renin angiotensin aldosterone system			
Tsai et al[119]	712 AF patients	rs5049/rs5051 (G217/G6) allele in AGT promoter gene	stroke
Stetskaia et al[121].	638 patients	rs4762 (T174M)	Stroke in females

** GWAS studies.

AF, atrial fibrillation, AGT; angiotensin; CES, cardioembolic stroke; LA, Left atrium; IS, ischemic stroke.

Open Peer Review | Bacteriology | Research Article

Bordetella pertussis exhibits genomic diversity within patients and laboratory culture

E. Bouda,[1] Y. Peng,[1] M. M. Williams,[1] P. K. Cassiday,[1] T. H. Skoff,[1] H. Ju,[1] H. Fueston,[1] D. Kurien,[2] D. Morales,[2] M. Cole,[1] E. A. Travanty,[3] D. Santoro,[4] P. M. Snippes Vagnone,[5] K. Mitchell,[6] M. Martinez,[7] Y. Unoarumhi,[8] M. Bentz,[8] S. Nobles,[8] J. Heuser,[2] M. L. Tondella,[1] L. C. Pawloski,[1] M. R. Weigand[1]

AUTHOR AFFILIATIONS See affiliation list on p. 10.

ABSTRACT Over the past decade, genomic characterization of *Bordetella pertussis* isolates recovered from US pertussis patients has unveiled noteworthy structural gene order variations. Whole-genome sequencing (WGS) shows that although *B. pertussis* exhibits little gene sequence variation, genomes from clinical isolates frequently differ in gene order through rearrangement between insertion sequence elements. To better understand rates of genome rearrangement and single nucleotide polymorphism (SNP) in *B. pertussis*, intra-patient genomic diversity was examined. Five states submitted, on average, five isolates per patient specimen following culture confirmation to the US Centers for Disease Control and Prevention for molecular characterization. Analysis of 149 patient specimen sets revealed only rare SNP variation, while isolate sets from 12 patients included genomic rearrangements that did not impact vaccine antigen production. To investigate the frequency and stability of such rearrangements during laboratory culture, replicate cultures of two pairs of isolates differing by duplication and inversion were subcultured for eight serial passages. WGS confirmed the initial presence of rare duplication mutations that became dominant in later passages, as well as the stable maintenance of a large, inverted genomic region during passage. These findings suggest that *B. pertussis* exhibits genomic diversity within a single clinical diagnostic specimen and acquires genomic variations during serial laboratory passages, indicative of bacterial genomic plasticity. Importantly, these rearrangements did not impact the frequency or distribution of SNPs. These results underscore the importance of minimizing the laboratory passaging of clinical isolates used for infectious disease surveillance.

IMPORTANCE The whooping cough-causing bacterium *Bordetella pertussis* can alter its genetic structure while conserving areas essential for vaccine efficacy. By examining *B. pertussis* from infected patients, we demonstrate the significance of reducing laboratory manipulation of bacteria to maintain reliable monitoring data from laboratory cultures. Public health officials can use these data to develop efficient disease control tactics and better understand *B. pertussis*'s adaptability. This study highlights the importance of immunizations and the necessity of thorough genetic surveillance in the fight against this chronic and avoidable respiratory illness.

KEYWORDS *Bordetella pertussis*, genomic rearrangements, genomic variations, genetic polymorphisms, whole genome sequencing, genome typing, serial passage, pulsed-field gel electrophoresis, genome assembly, genome duplication

Pertussis (whooping cough), a debilitating respiratory disease caused by the bacterium *Bordetella pertussis*, remains a frequently reported vaccine-preventable disease worldwide. Beginning in the 1990s, acellular pertussis (aP) vaccine formulations containing pertussis toxin (Pt), filamentous hemagglutinin (Fha), and pertactin (Prn), with

Editor Marnix Medema, Wageningen University, Wageningen, Netherlands

Peer Reviewer Andrew Preston, University of Bath, Bath, United Kingdom

Address correspondence to M. R. Weigand, mweigand@cdc.gov.

The authors declare no conflict of interest.

Received 27 December 2024
Accepted 27 April 2025
Published 17 June 2025

This is a work of the U.S. Government and is not subject to copyright protection in the United States. Foreign copyrights may apply.

Downloaded from https://journals.asm.org/journal/msystems on 15 October 2025 by 24.117.150.43.

or without fimbriae types 2 and 3 (Fim2/3), replaced whole-cell pertussis (wP) vaccines in many industrialized countries. Following the switch, an increased incidence of pertussis was reported in the USA and elsewhere (1–3) until the start of the coronavirus disease 2019 pandemic (4) despite widespread vaccination coverage (5). Multiple factors likely explain the resurgence of pertussis cases, including more rapidly waning immunity conferred by aP vaccination compared to wP vaccine formulations (5, 6), as well as improvements in diagnostic testing and reporting (7).

The bacterium appears to have changed following the switch to aP vaccine formulations. Specifically, the emergence of new alleles *ptxP3, prn2,* and *fim3B,* as well as frequent gene disruption causing Prn deficiency (8–16), has been suggested to be adaptive responses to aP vaccine-induced immune pressure. Despite these mutations, traditional molecular typing methods have described *B. pertussis* as genetically monomorphic due to the paucity of single nucleotide polymorphisms (SNPs) among clinical isolates (17, 18). Somewhat paradoxically, pulsed-field gel electrophoresis (PFGE) often revealed additional diversity, including among clinical isolates within epidemics, though the underlying mutations causing such variations remained unverified (19–23).

Through recent advances in whole-genome sequencing (WGS), long-read technology allowed profound sequence-based characterization of circulating *B. pertussis*. Indeed, the alignment of several complete genome assemblies from clinical isolates revealed that current circulating isolates exhibit considerable diversity of gene order arrangements, both compared to laboratory reference strains and each other (24, 25). These gene order differences appeared as large duplications and inversions facilitated by insertion sequence (IS) elements, notably IS*481* (24, 25). However, these studies have been limited to characterizing genetic differences between representative, single colonies from multiple patients (8, 21, 25–29), and little is known about the frequency or stability of genomic rearrangements during infection of human hosts or laboratory culture.

The current study combined traditional molecular typing and WGS to examine the frequency of *B. pertussis* genetic variation in infected patients at the single time point of diagnosis. It also examined the impact of subsequent laboratory subculturing on genome stability. The findings presented in this study demonstrate *B. pertussis* genomic rearrangement plasticity during growth in infected patients and laboratory culture. Therefore, limiting laboratory passage during genomic characterization is crucial for accurate genomic surveillance.

RESULTS

High nucleotide sequence conservation among replicate isolates from individual patients

The current study analyzed 705 isolates of *B. pertussis* recovered in sets of 2 to 10 colonies per patient during culture confirmation of positive nasopharyngeal specimens from 149 patients (Fig. 1A and B; Table S1). Sequence variation within the bacterial population infecting individual patients was evaluated by comparing genome-wide SNP profiles between isolates in each patient set and summarized using the maximum pairwise SNP distance observed between isolates in the set. As anticipated, the observed number of SNPs within sets was relatively low. Among the 149 patient isolate sets, 115 (77%) exhibited no SNPs, 16 (11%) displayed one to three SNP differences, and only 18 (12%) presented more than three SNPs within sets (Fig. 1C). This observation of high genetic similarity within patients aligns with the current understanding of nucleotide sequence conservation among circulating *B. pertussis* (18, 30).

Chromosome rearrangements detected among replicate isolates from individual patients

Genetic variation among isolates in each set was investigated through PFGE and WGS. First, PFGE profiles of the 705 isolates studied (Table S2) revealed that sets from 12 patients exhibited variation, observable as the appearance of additional bands (Fig. 2A)

Downloaded from https://journals.asm.org/journal/msystems on 15 October 2025 by 24.117.150.43.

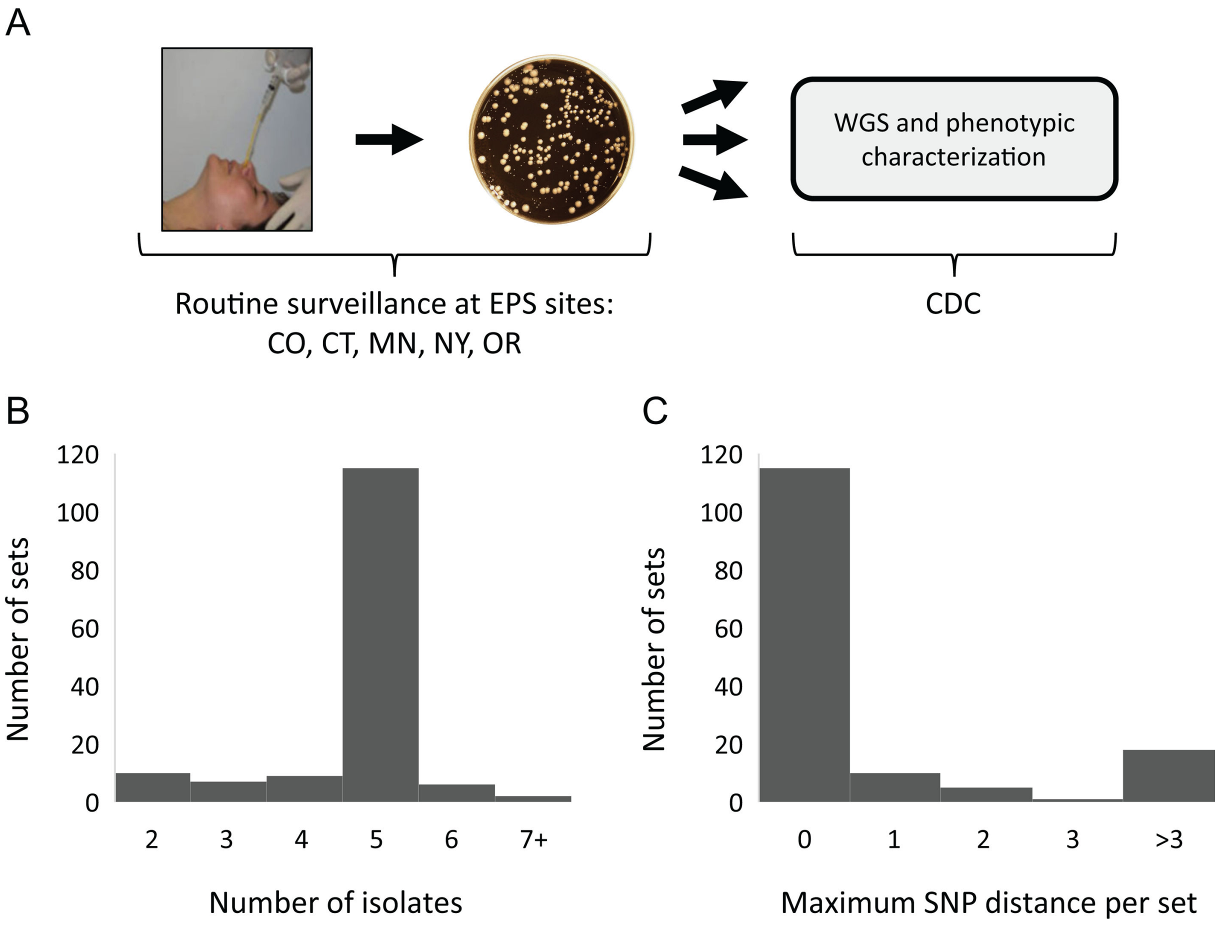


FIG 1 Genome sequence variation among isolates from the same patient. (A) Specimen collection and processing (image courtesy of CDC Public Health Image Library). (B) Histogram depicting the number of isolates per set. Eighty-nine percent of sets included at least four isolates. (C) Rare SNPs within isolate sets from 149 patients analyzed. Seventy-seven percent of sets had zero SNP differences within sets; 11% of sets had one to three SNP differences; and only 12% had more than three SNPs within sets. Reported values represent maximum pairwise SNP distances.

or reshuffling of bands (Fig. 2C) in profiles of one or two isolates that differed from others in those patient sets. Short-read and long-read sequencing platforms provided additional resolution to investigate potential chromosomal variations captured by the PFGE profile differences among select isolates in patient sets (11, 15, 19, 20, 32, 68, 73, 75, 93, 113, 131, 150). Whole-genome alignment of complete *de novo* assemblies was used to identify evidence of large chromosomal rearrangements, such as duplications, inversions, or translocations (Fig. S1A, B and S2). Such alignments revealed inversion differences in seven patient sets. For example, in patient set 20 (Table S2), genomes of isolates J817 and J818 exhibited structure CDC300, while J816, J819, and J820 had structure CDC237 (Fig. 2C). Structures CDC237 and CDC300, named according to their associated PFGE profiles, differ by a single large inverted genomic segment of approximately 1 Mbp (Fig. 2D), as described in previous studies (25, 26).

In a similar manner, mapping short Illumina reads to a reference genome was used to investigate potential duplications, observable as localized increases in coverage depth in five patient sets at distinct genomic regions (Fig. S1A). For instance, sequencing reads from three isolates in patient set 15 (J763, J764, J765) exhibited even coverage across the 4.1 Mbp reference genome. However, reads from J766 and J767 of that same patient set 15 each yielded regional increases in sequencing coverage at positions 1.69–1.96 Mbp and 775–855 kbp, respectively (Fig. 2B). The observed coverage peak in J767 corresponded to approximately twice the genome-wide average, suggesting the 80 kbp region was now present in two copies, similar to duplication events previously observed in

Downloaded from https://journals.asm.org/journal/msystems on 15 October 2025 by 24.117.150.43.

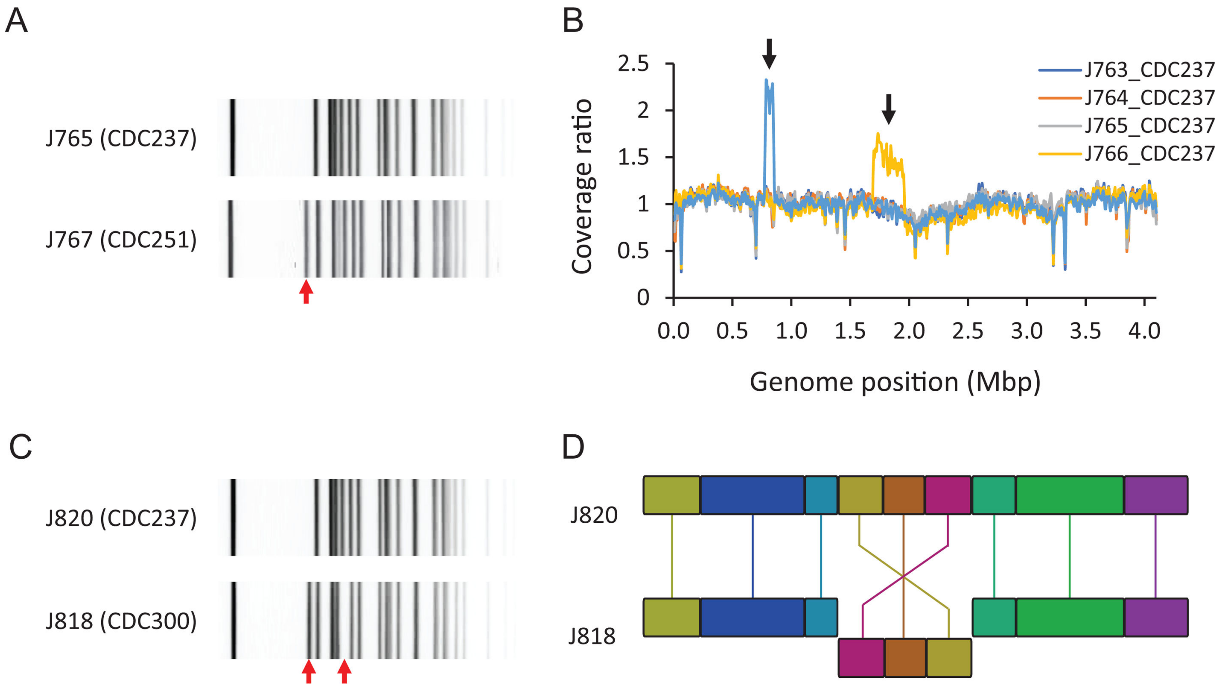


FIG 2 Genome variations among isolates from the same patient. (A) The PFGE profile of J767 contained an additional band (bottom and red arrow) that differentiated it from the profile of CDC237 and the rest of that patient's set (J763, J764, J765, and J766). (B) WGS identified the variation in J767 as a regional duplication (blue arrow). An increase in coverage is also observed with J766 (yellow arrow) but did not reach 2× coverage. (C) The banding pattern of the PFGE profiles of a patient set from CO differed by two bands (bottom and red arrows), consistent with CDC structures 237 and 300. (D) WGS identified the variation detected in panel C as a regional genomic inversion. In panel B, genome coordinates are indicated in the *x*-axis; the *y*-axis indicates the number of times each position on the genome coordinate is observed.

other studies (24). By contrast, the observed coverage peak in J766 did not reach 2× the genome average. Furthermore, PacBio long-read sequencing and assembly confirmed duplication of the genomic region corresponding to the coverage peak in J767 but not J766.

Stability of genomic inversions and emergence of duplications during laboratory passage

To assess the frequency and stability of the genetic differences observed within patient sets, genomic DNA was sequenced from serial laboratory passages at multiple intermediate timepoints. Two representative isolate pairs, J766/J767 from patient set 15 and J818/J820 from patient set 20, were selected based on their within-set genomic variation of duplication and inversion, respectively. Five replicate lineages of each were maintained on Regan-Lowe charcoal agar without cephalexin for eight serial passages. To track duplications, shotgun Illumina sequencing reads of passages 1, 4, and 8 were mapped to a reference *de novo* assembly (GenBank: CP032766.1) and confirmed that the observed duplication in J767 remained stable throughout the eight serial passages (Fig. 3A). However, in lineages of J766, a new duplicated region emerged (position 2,525,000–2,640,000), first detected at passage 2 and steadily increasing in abundance through passage 8 (Fig. 3B). Duplication within the genome of J766 became detectable within two passages of subculturing, perhaps because it existed undetected within a small subpopulation and rapidly expanded during serial passage. The hypothesis that the duplication existed in low frequency before passage is corroborated by the fact that all five lineages accumulated the same duplication throughout the eight passages.

Stability of a large inversion was monitored by mapping long PacBio sequencing reads from replicate passages 1 and 8 of J818/J820 to a reference *de novo* assembly

Downloaded from https://journals.asm.org/journal/msystems on 15 October 2025 by 24.117.150.43.

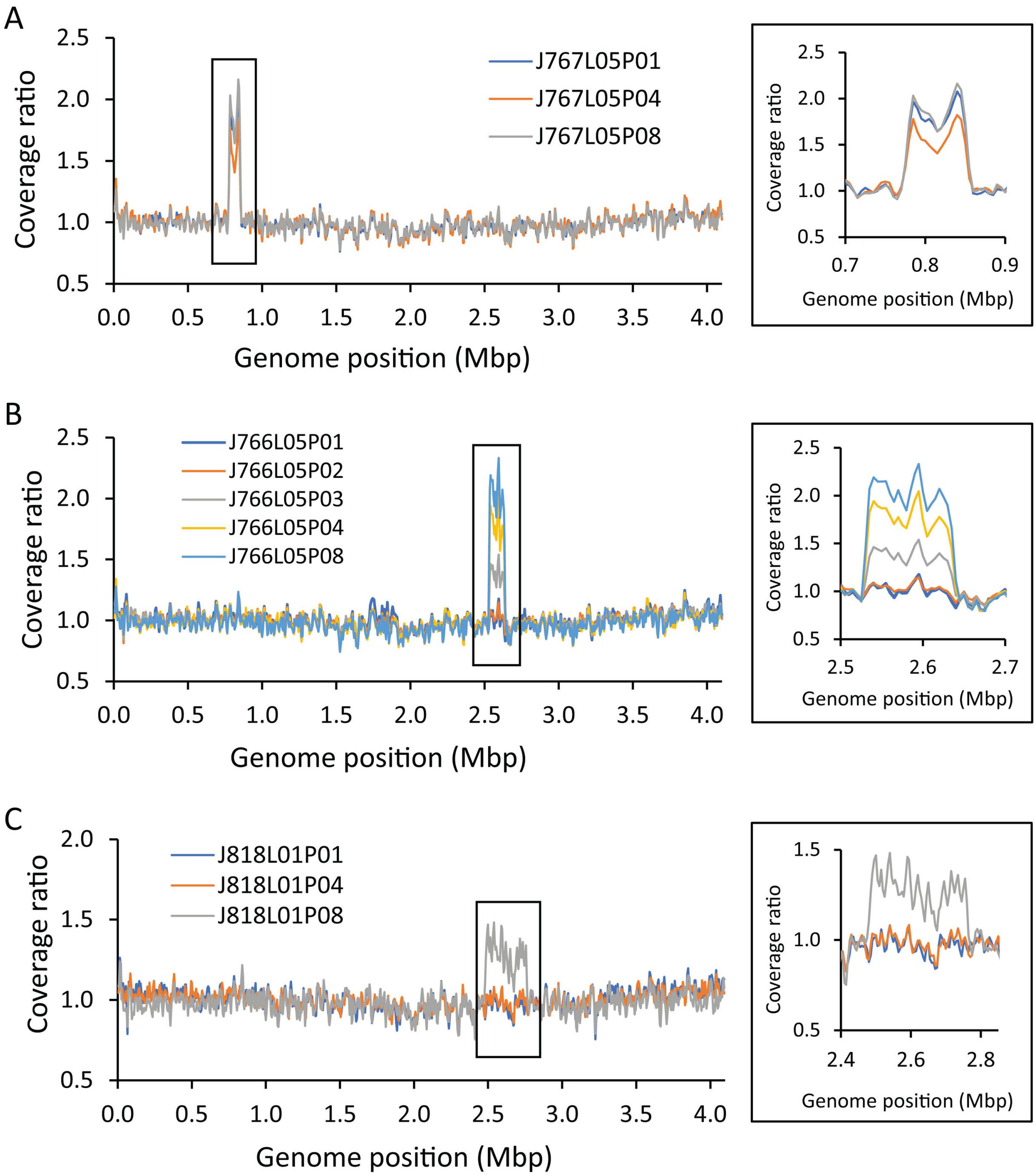


FIG 3 Dynamics of genome rearrangement during laboratory culture. The plots compare coverages for different passages (P01, P02, P03…) of isolates J767 (A), J766 (B), and J818 (C) using the same lineage per isolate. The plots show increases in depth in different regions in all three isolates. Zooming in at the increased coverage (inset, top right of each panel) confirmed the presence of a duplication stably maintained from passage 1 of J767 (A) and the progressive accumulation of a subpopulation of isolate J766 containing a duplication (B) distinct from that of J767. (C) WGS analysis of the serial passage of the pair of isolates J818/J820 initially investigated for a regional inversion revealed a regional duplication not previously observed at passage 8. Genome coordinates are in kilobases (kb) as indicated in the *x*-axis. The *y*-axis indicates the number of times each position on the genome coordinate is observed.

(GenBank: CP043236). The primary difference between the genomes of J818 and J820, an inverted segment of approximately 1 Mbp, appeared stable throughout the eight serial passages (Fig. 2D), and no evidence of inversion was observed during the serial passage in any of the replicate lineages. However, a new duplication event emerged in one of the five replicates during the final passage (Fig. 3C) for both J818 and J820. Mapping Illumina

Downloaded from https://journals.asm.org/journal/msystems on 15 October 2025 by 24.117.150.43.

short-read sequences at passages 1, 4, and 8 suggested that the duplication event likely occurred between passages 5 and 8—similar timing with changes observed in previous subculturing experiments analyzed by PFGE (31).

Frequent Prn and Fim2 deficiency among isolates

Characterizing the 705 isolates from the 149 patient sets included a protein production screening assay and multilocus sequence typing to look for variations in vaccine immunogens within each patient set (Table S1). Most isolates did not produce detectable Prn or Fim2, with deficiency observed in 132 (89%) and 139 (93%) sets, respectively. WGS analysis showed that 87% of mutations underlying Prn deficiency were IS*481* insertion within *prn* (Table S2), consistent with findings in previous studies (14). By comparison, only 7% of sets (54, 55, 70, 73, 80, 88, 98, 100, 121) lacked detectable Fim3 (Table S1). Interestingly, two sets (131, 132) contained one isolate each deficient in Fha, and only set 134 contained one isolate deficient in Pt (Table S1). Allele typing for *fim3, prn, ptxA,* and *ptxP* showed that isolates within sets consistently displayed the same sequence type, and no variations were observed (Table S1).

DISCUSSION

Previous studies have shown that circulating *B. pertussis* diverged from vaccine reference strains through a combination of genome sequence and gene order changes (1, 25). For an in-depth analysis of these dynamics during infection, this study utilized traditional molecular assays and WGS to assess *B. pertussis* genetic variation within infected patients and during subsequent laboratory passage, respectively. The analyses described herein revealed that although bacterial populations within patients may exhibit minor genetic variation, these rarely result in detectable differences in aP vaccine immunogen production. However, the genome sequence and structural dynamics within patient diagnostic specimens highlight critical implications for accurate interpretation of molecular epidemiology and surveillance using cultured isolates.

While *B. pertussis* has historically been regarded as a monomorphic organism (18), a growing number of reports from comparative genomic analyses have indicated notable heterogeneity among clinical isolates (24, 25). Detecting genomic variation by molecular typing and WGS among replicate isolates from a single patient collected at a single time point here now adds another layer of complexity to our understanding of *B. pertussis* evolution. Specifically, such observation suggests that genomic rearrangements likely occur during bacterial proliferation within infected patients and are not artifacts of subsequent clinical laboratory manipulation. These structural changes, both inversions and duplications, underscore the remarkable genomic plasticity of *B. pertussis,* primarily driven by high-copy numbers of IS*481* (23).

Previous studies of repeated subculture with *B. pertussis* have yielded mixed findings. One study reported that sporadic *B. pertussis* PFGE profile changes can occur during serial passaging (31), and another observed mutation leading to reduced expression of type III secretion (32–34), while another reported no genomic alterations after 11 consecutive passages (25). The passage results here suggest that large genomic inversions, but not duplications, remain stable during repeated subculture, suggesting that perhaps previously reported changes in PFGE profile reflected duplications only. While duplications appeared frequently during passage in this study, the question remains whether this rearrangement was present at low frequency within the initial population, becoming observable only through repeated culturing. However, even if true, the hypothesized increasing frequency observed in subsequent passages suggests the duplications were selectively advantageous during laboratory culture, allowing those harboring it to outcompete population members without it. The phylogenetic distribution of duplications detected in publicly available *B. pertussis* genomic data reported previously highlighted multiple homoplasies, also consistent with potential adaptation to laboratory culture (24). Perhaps direct sampling from patients at multiple time points

Downloaded from https://journals.asm.org/journal/msystems on 15 October 2025 by 24.117.150.43.

throughout infection may be essential to unraveling lingering questions regarding the delicate interplay between genomic stability and adaptability of *B. pertussis*.

Applications of high-throughput WGS in public health have transformed molecular epidemiology and surveillance of microbial pathogens with added resolution (35). However, the implementation of WGS for tracking pertussis clusters has been challenged, in part, by the pathogen's very low mutation rate, leading some to suggest that distance thresholds as low as two SNPs may define linked cases (28, 36, 37). Observed SNP differences among replicate isolates here could have implications for genomic epidemiology as case clustering results may differ depending on which isolate is selected for analysis. Similarly, although the use of PFGE for molecular characterization has declined, many previous studies noted that this assay often indicated higher levels of diversity compared to other methods, such as allele typing (19, 31). The results here suggest that PFGE may have overestimated phylogenetic diversity within pertussis epidemics due to the dynamic nature of duplications in particular and, to a lesser extent, inversions. Perhaps culture-independent WGS (38) directly from nasopharyngeal specimens, as well as controlled human infection model experiments, can shed more light on *B. pertussis* genomic diversity within single infections to both detect and account for such variation.

It is important to acknowledge the limitations of this study. While the data presented here provide valuable insights, there were few isolates per patient, limiting the detection of genetic variations that may have been present at a low frequency. Additionally, to minimize the burden on participating surveillance sites, this study focused on replicate isolates collected at a single time point (e.g., clinical diagnosis). In contrast, longitudinal studies could provide deeper insights into genetic diversification throughout the course of infection. Critically, both limitations suggest that the present study could not capture the full extent of variation present within the bacterial population infecting individual patients, and the number of genomic rearrangements observed here is likely underestimated. Similarly, the very low SNP distances observed within sets here prohibit confidently determining whether such mutations emerged during infection or represent colonization from diverse sources within a transmission network. Numerous studies have established that bacterial populations accumulate genetic diversity during transmission and even within single colonies on agar media, thus potentially colonizing patients with heterogeneous genotypes (39–41). However, the present study underscores pertussis-specific challenges presented by the pathogen's relatively low mutation rate, limiting genomic epidemiology and surveillance.

In conclusion, integrating genomic, molecular, and laboratory passage data in this study increased our understanding of *B. pertussis* genome plasticity. These findings underscore the importance of ongoing surveillance and highlight the potential challenges posed by genetic diversity in the context of pertussis control. As *B. pertussis* evolves, adaptive public health programs will be essential to effectively combat this persistent respiratory pathogen. It is important that future research continues to monitor genetic diversity, including genome rearrangements, to better understand mechanisms driving *B. pertussis* evolution that may affect pertussis control.

MATERIALS AND METHODS

Isolate collection

All *B. pertussis* isolates used in this study were collected between 2016 and 2022 through enhanced pertussis surveillance (EPS) (7) at five participating sites, each providing isolates from a convenience sample of patients that were not representative of all captured cases in each region: Colorado ($n = 42$), Connecticut ($n = 4$), Minnesota ($n = 67$), New York ($n = 9$), and Oregon ($n = 29$). All patients met the Council of State and Territorial Epidemiologists (CSTE) pertussis case definition except for two patients from Oregon. Nasopharyngeal specimens positive for pertussis were confirmed by culture. Multiple colonies were recovered from each patient specimen, harvesting colonies with maximal

Downloaded from https://journals.asm.org/journal/msystems on 15 October 2025 by 24.117.150.43.

morphological differences where possible and submitted as a “set” of 2 to 10 replicate isolates (average = 5) to CDC for characterization. Colony morphological differences were not systematically tracked during subsequent culturing at CDC. In total, 705 isolates from 149 patients were included (Table S1). The two patients from Oregon, having sets 73 and 98 (Table S2), were inadvertently characterized in these studies as EPS sets and counted among the 149 patients; only these two patients did not meet the CSTE pertussis case definition or fall within the EPS catchment area but suited the laboratory characterization parameters. Patient demographics can be found in Table S3.

Bacterial cell culture and DNA extraction

Isolates were cultured on Regan-Lowe charcoal agar media without cephalexin for 72 hours at 37°C and harvested as previously described (28) for DNA extraction. Genomic DNA extraction was performed using the Maxwell RSC Instrument and the Maxwell RSC Whole Blood DNA Kit (Promega, AS1520). All DNA samples were quantified and qualified using a NanoDrop 8000 Spectrophotometer (Thermo Fisher Scientific Inc., Wilmington, DE) and a Qubit Fluorometer 3.0 with the Qubit dsDNA Broad Range Assay Kit (Life Technologies, Waltham, MA).

Pulsed-field gel electrophoresis

PFGE was performed on all 705 isolates at CDC, as previously described (42). Briefly, extracted DNA was embedded in agarose plugs and digested with the restriction enzyme XbaI. Agarose gels (1%) were prepared, and the restricted DNA plugs were placed onto the gel, and electrophoresis was conducted in 0.5× Tris-borate-EDTA buffer at 6 V/cm for 18 hours at 14°C with a ramped switch time of 2.2 to 35 s. Gels were stained with GelRed, and bands were visualized using UV light. Captured digital TIFF images were analyzed using BioNumerics software, version 5.01 (Applied Maths, Inc., Austin, TX), and profiles were compared to the CDC database and assigned based on bands in the range of 125 to 525 kb.

Whole-genome sequencing and assembly

Whole-genome shotgun sequencing of isolates was performed using a combination of the PacBio Sequel II (Pacific Biosciences, Menlo Park, CA) and Illumina MiSeq (Illumina, San Diego, CA) platforms as described previously (26). Briefly, Illumina paired-end libraries were prepared with the NEBNext Ultra II DNA Library Prep Kit (New England Biolabs, Ipswich, MA) and sequenced on a MiSeq using v.2 Kit (Illumina, San Diego, CA). The Illumina raw reads were checked for quality using FastQC v.0.11.5 (43), and then trimmed and filtered with Cutadapt v.2.3 (44). Libraries were prepared for PacBio sequencing using SMRTbell template prep kit 2.0, SMRTbell Enzyme Clean Up Kit 2.0, Sequel Bind Kit 2.2, and Sequel II Sequencing Kit 2.0. A library size selection was performed using BluePippin with 0.75% agarose gels. *De novo* assembly was performed using FLYE (v.2.4.2) with default settings (45). The resulting consensus sequences were manually checked for circularity using Gepard (v.1.30) (46) and then reordered to match the start of Tohama I (GenBank accession no. NC_002929). Assemblies were further polished by mapping Illumina reads using the CLC Genomics Workbench (v.23; CLC bio, Boston, MA). Assemblies were annotated using the NCBI Prokaryotic Genome Annotation Pipeline.

Sequence variation within sets

SNPs were determined by mapping trimmed Illumina short reads to a closed genome assembly from an isolate of the same set using Snippy v.4.3.8 (47) with default settings. Pairwise SNP distances within each set were calculated from a core genome alignment using SNP-Dist v.4.0 (48) and summarized using the maximum pairwise distance. Large duplications were identified by mapping trimmed Illumina short reads to a closed genome assembly from an isolate of the set using bowtie v.2.2.9 (49) and inspecting

Downloaded from https://journals.asm.org/journal/msystems on 15 October 2025 by 24.117.150.43.

the coverage depth calculated with samtools v.1.8 (50) plotted across a sliding 10 kbp window. Inversions and other rearrangements were determined by pairwise alignment of assembled PacBio genomes using MUMmer v.4 (51).

Protein production screening

Isolates were cultured on Regan-Lowe agar without cephalexin for 72 hours at 37°C. Bacterial cells were collected and suspended in 0.01 M phosphate-buffered saline (PBS) to a turbidity of 1.0, then measured using a Dade Behring MicroScan Turbidity Meter (Siemens Global, Munich, Germany). Cells were pelleted by centrifugation at 10,000 × *g* and resuspended in Xtractor lysis buffer (Clontech, CA, USA) containing 25 U of Benzonase (Sigma-Aldrich, MO, USA) at a 1:1 ratio (vol/vol). The mixture was vortexed gently for 10 min at room temperature and centrifuged, and lysis buffer supernatant was discarded. Cell lysate pellets were resuspended in PBS for analysis, as previously described (52, 53).

The electrochemiluminescent antibody capture assay was performed using the Sector S6000 platform (Meso-Scale Discovery, Maryland, USA). Purified Prn, Fha, Pt, Fim2, and Fim3 at 1 µg/mL (GlaxoSmithKline, UK) were used as positive controls for the respective immunogens. Respective 96-well plates per immunogen (Meso-Scale Discovery) were coated with 30 µL of 1/1,500 working dilution (wd) of rabbit anti-Prn polyclonal antibody (Alpha Diagnostics, TX), capture sheep Fha anti-serum at 1/33,000 wd, capture sheep Pt anti-serum at 1/10,000 wd, capture human anti-*B. pertussis* serum at 1/1,500 wd, and capture sheep Fim3 anti-serum at 1/20,000 wd (NIBSC, UK), sealed, and incubated overnight at 4°C. The next day, plates were washed three times with at least 200 µL of wash buffer containing PBS and 0.05% Tween (Fisher Scientific, PA, USA). Blocking buffer (150 µL) containing Blocker A (1.0 g, Meso-Scale Discovery), milk powder (2.5 g, Fisher Scientific, PA, USA), goat serum (1.0 mL, Sigma-Aldrich, MO, USA), and PBS (46 mL) were added to the plates, sealed, and incubated on a plate shaker at room temperature (RT) for at least 30 min. Blocking buffer was replaced with 25 µL/well of cell lysates and incubated at RT for 1 hour. Plates were washed 6× with at least 200 µL of wash buffer and incubated for 45 min at RT or 4°C overnight with 25 µL of sheep anti-Prn detection antibody at 1/4,000 wd, mouse anti-Fha 1/5000 wd, mouse anti-Pt subunit 1 at 1/1,000 wd, mouse anti-Fim2 at 1/1,500 wd, and mouse anti-Fim3 at 1/20,000 wd, respectively (NIBSC, UK). Plates were washed three times with at least 200 µL of wash buffer and incubated for 45 min at RT or 4°C overnight with 25 µL/well of donkey anti-sheep Sulfo-tag secondary antibody at 1 µg/mL for Prn and goat anti-mouse Sulfo-tag at 1 µg/mL for Fha, Pt, Fim2, and Fim3 (Meso_Scale Discovery). Plates were washed three times with at least 200 µL of wash buffer, then replaced with 150 µL 2× read buffer. Luminescence was measured with a Sector Reader (Meso-Scale Discovery). A twofold dilution series with 2 µg/mL purified Prn, Fha, Pt, Fim2, or Fim3 was performed and fit with a four-parameter logistic regression using the instrument software to confirm the absence of nonspecific cross-reactivity. Isolates were qualitatively defined as either Prn-, Fha-, Pt-, Fim2-, or Fim3-producing or -deficient based on their signal intensity compared to positive and negative controls measured on the same plate to account for variability between plates (52). The assay was repeated for confirmation for any isolate producing <50% signal intensity compared to the control.

Serial laboratory passage

Two pairs of isolates (J766/J767 and J818/J820) were selected for serial laboratory passaging at CDC to assess the stability of genomic rearrangements. Bacterial cell culture was performed as described above, and five replicate lineages of each isolate were transferred to fresh Regan-Lowe charcoal agar plates without cephalexin (RL) every 72 hours. Each lineage was started by rubbing a single bead from a frozen stock on RL. After 72 hours, plates were checked for contamination, then half of the growth was harvested using a cotton swab and resuspended in 600 µL of saline for DNA extraction, library prep, and sequencing as described above. One-third of the plate growth was used to prepare

Downloaded from https://journals.asm.org/journal/msystems on 15 October 2025 by 24.117.150.43.

cryobeads (Microbank, Pro-Lab Diagnostics, Round Rock, TX) for long-term storage at −70°C, according to the manufacturer's instructions. The remaining growth was used to inoculate a fresh RL plate. This process was repeated eight times. Unique identifiers were assigned to track each isolate's replicate lineages (L) and passages (P). For example, "J766L01P01" indicates J766 lineage 1 and passage 1.

The emergence of SNPs and large duplications during passages was investigated by mapping Illumina reads to genomes assembled before passage in a similar manner as described above. Potential genome rearrangements occurring during the passage of J818 and J820 were investigated by mapping PacBio reads to their previous genome assemblies using Minimap2 (54, 55) and calling structural variants with Sniffles (56).

ACKNOWLEDGMENTS

We thank the Enhanced Pertussis Surveillance Network, Amy Rubis (CDC), Mili Sheth (CDC), the CDC Genome Sequencing Laboratory, the CDC Laboratory Leadership Service Program, and the United States Public Health Service (USPHS) Commissioned Corps for their support.

This work was made possible through support from CDC's Advanced Molecular Detection (AMD) program.

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention. Use of trade names and commercial sources is for identification only and does not imply endorsement by the Centers for Disease Control and Prevention, the Public Health Service, or the US Department of Health and Human Services.

AUTHOR AFFILIATIONS

[1]Division of Bacterial Diseases, National Center for Immunization and Respiratory Diseases, Centers for Disease Control and Prevention, Atlanta, Georgia, USA
[2]ASRT, Inc., Atlanta, Georgia, USA
[3]Colorado State Public Health Laboratory, Denver, Colorado, USA
[4]Public Health Laboratory, Connecticut Department of Public Health, Rocky Hill, Connecticut, USA
[5]Public Health Laboratory Division, Minnesota Department of Health, Saint Paul, Minnesota, USA,
[6]New York State Public Health Laboratory, New York, New York, USA
[7]Oregon State Public Health Laboratory, Hillsboro, Oregon, USA
[8]Division of Core Laboratory Services and Response, Office of Laboratory Systems and Response, Centers for Disease Control and Prevention, Atlanta, Georgia, USA

PRESENT ADDRESS

E. Bouda, Office of the Director, National Center for HIV, Viral Hepatitis, STD and TB Prevention, Centers for Disease Control and Prevention, Atlanta, Georgia, USA
M. M. Williams, Division of Healthcare Quality Promotion, National Center for Emerging and Zoonotic Infectious Diseases, Centers for Disease Control and Prevention, Atlanta, Georgia, USA
M. R. Weigand, Division of Core Laboratory Services and Response, Office of Laboratory Systems and Response, Centers for Disease Control and Prevention, Atlanta, Georgia, USA

AUTHOR ORCIDs

M. Cole http://orcid.org/0000-0002-9205-9055
Y. Unoarumhi http://orcid.org/0000-0001-9088-0360
M. R. Weigand http://orcid.org/0000-0002-7278-0160

Downloaded from https://journals.asm.org/journal/msystems on 15 October 2025 by 24.117.150.43.

AUTHOR CONTRIBUTIONS

E. Bouda, Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Validation, Writing – original draft, Writing – review and editing | Y. Peng, Conceptualization, Data curation, Formal analysis | M. M. Williams, Conceptualization, Methodology, Writing – review and editing | P. K. Cassiday, Data curation, Formal analysis, Methodology, Writing – review and editing | T. H. Skoff, Conceptualization, Investigation, Writing – review and editing | H. Ju, Data curation, Writing – review and editing | H. Fueston, Data curation, Writing – review and editing | D. Kurien, Data curation, Writing – review and editing | D. Morales, Writing – review and editing | M. Cole, Data curation, Writing – review and editing | E. A. Travanty, Data curation, Investigation, Writing – review and editing | D. Santoro, Data curation, Writing – review and editing | P. M. Snippes Vagnone, Data curation, Writing – review and editing | K. Mitchell, Data curation, Writing – review and editing | M. Martinez, Data curation, Writing – review and editing | Y. Unoarumhi, Data curation, Formal analysis | M. Bentz, Data curation, Visualization | S. Nobles, Data curation | J. Heuser, Data curation, Visualization | M. L. Tondella, Conceptualization, Supervision, Writing – review and editing | L. C. Pawloski, Conceptualization, Data curation, Methodology, Supervision, Writing – review and editing | M. R. Weigand, Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Supervision, Validation, Visualization, Writing – review and editing

DATA AVAILABILITY

Surveillance isolate whole-genome sequence data are available from NCBI under BioProject accession number PRJNA279196 and sequence data from serial laboratory passage experiments are available under PRJNA1258996, as listed in Tables S2 and S4, respectively.

ETHICS APPROVAL

This activity was reviewed by CDC, was deemed not research, and was conducted consistent with applicable federal law and CDC policy. See e.g., 45 C.F.R. part 46, 21 C.F.R. part 56; 42 U.S.C. §241(d); 5 U.S.C. §552 a; 44 U.S.C. §3501 et seq.

ADDITIONAL FILES

The following material is available online.

Supplemental Material

Supplemental figures (mSystems01717-24-s0001.pdf). Figures S1 and S2.
Table S1 (mSystems01717-24-s0002.xlsx). The 149 patient sets.
Table S2 (mSystems01717-24-s0003.xlsx). Individual isolates with corresponding PFGE, Pertactin mutation data, and accession numbers.
Table S3 (mSystems01717-24-s0004.xlsx). Patient demographics.
Table S4 (mSystems01717-24-s0005.xlsx). Lineage accession numbers.

Open Peer Review

PEER REVIEW HISTORY (review-history.pdf). An accounting of the reviewer comments and feedback.

REFERENCES

1. Bart MJ, Harris SR, Advani A, Arakawa Y, Bottero D, Bouchez V, Cassiday PK, Chiang C-S, Dalby T, Fry NK, et al. 2014. Global population structure and evolution of *Bordetella pertussis* and their relationship with vaccination. mBio 5:e01074. https://doi.org/10.1128/mBio.01074-14
2. Clark TA. 2014. Changing pertussis epidemiology: everything old is new again. J Infect Dis 209:978–981. https://doi.org/10.1093/infdis/jiu001
3. Tan T, Dalby T, Forsyth K, Halperin SA, Heininger U, Hozbor D, Plotkin S, Ulloa-Gutierrez R, Wirsing von König CH. 2015. Pertussis across the

Downloaded from https://journals.asm.org/journal/msystems on 15 October 2025 by 24.117.150.43.

globe: recent epidemiologic trends from 2000 to 2013. Pediatr Infect Dis J 34:e222–e232. https://doi.org/10.1097/INF.0000000000000795

4. Pertussis surveillance and trends. 2024. US Centers for Disease Control and Prevention. Available from: https://www.cdc.gov/pertussis/php/surveillance/index.html
5. Szwejser-Zawislak E, Wilk MM, Piszczek P, Krawczyk J, Wilczyńska D, Hozbor D. 2022. Evaluation of whole-cell and acellular pertussis vaccines in the context of long-term herd immunity. Vaccines (Basel) 11:1. https://doi.org/10.3390/vaccines11010001
6. Noel G, Borand L, Leng C, Keang C, Botr C, Dim B, Kerleguer A, Peng YS, Sreng N, Ork V, Ait-Ahmed M, Guiso N, Taieb F. 2021. Circulation of *Bordetella pertussis* in vaccinated Cambodian children: a transversal serological study. Int J Infect Dis 106:134–139. https://doi.org/10.1016/j.ijid.2021.03.054
7. Skoff TH, Baumbach J, Cieslak PR. 2015. Tracking pertussis and evaluating control measures through enhanced pertussis surveillance, emerging infections program, United States. Emerg Infect Dis 21:1568–1573. https://doi.org/10.3201/eid2109.150023
8. Weigand MR, Peng Y, Cassiday PK, Loparev VN, Johnson T, Juieng P, Nazarian EJ, Weening K, Tondella ML, Williams MM. 2017. Complete genome sequences of *Bordetella pertussis* isolates with novel pertactin-deficient deletions. Genome Announc 5:37. https://doi.org/10.1128/genomeA.00973-17
9. Barkoff A-M, Mertsola J, Guillot S, Guiso N, Berbers G, He Q. 2012. Appearance of *Bordetella pertussis* strains not expressing the vaccine antigen pertactin in Finland. Clin Vaccine Immunol 19:1703–1704. https://doi.org/10.1128/CVI.00367-12
10. Bouchez V, Brun D, Cantinelli T, Dore G, Njamkepo E, Guiso N. 2009. First report and detailed characterization of *B. pertussis* isolates not expressing pertussis toxin or pertactin. Vaccine (Auckl) 27:6034–6041. https://doi.org/10.1016/j.vaccine.2009.07.074
11. Hegerle N, Paris A-S, Brun D, Dore G, Njamkepo E, Guillot S, Guiso N. 2012. Evolution of French *Bordetella pertussis* and *Bordetella parapertussis* isolates: increase of *Bordetellae* not expressing pertactin. Clin Microbiol Infect 18:E340–E346. https://doi.org/10.1111/j.1469-0691.2012.03925.x
12. Otsuka N, Han H-J, Toyoizumi-Ajisaka H, Nakamura Y, Arakawa Y, Shibayama K, Kamachi K. 2012. Prevalence and genetic characterization of pertactin-deficient *Bordetella pertussis* in Japan. PLoS One 7:e31985. https://doi.org/10.1371/journal.pone.0031985
13. Queenan AM, Cassiday PK, Evangelista A. 2013. Pertactin-negative variants of *Bordetella pertussis* in the United States. N Engl J Med 368:583–584. https://doi.org/10.1056/NEJMc1209369
14. Pawloski LC, Queenan AM, Cassiday PK, Lynch AS, Harrison MJ, Shang W, Williams MM, Bowden KE, Burgos-Rivera B, Qin X, Messonnier N, Tondella ML. 2014. Prevalence and molecular characterization of pertactin-deficient *Bordetella pertussis* in the United States. Clin Vaccine Immunol 21:119–125. https://doi.org/10.1128/CVI.00717-13
15. Bottero D, Gaillard ME, Basile LA, Fritz M, Hozbor DF. 2012. Genotypic and phenotypic characterization of *Bordetella pertussis* strains used in different vaccine formulations in Latin America. J Appl Microbiol 112:1266–1276. https://doi.org/10.1111/j.1365-2672.2012.05299.x
16. Schmidtke AJ, Boney KO, Martin SW, Skoff TH, Tondella ML, Tatti KM. 2012. Population diversity among *Bordetella pertussis* isolates, United States, 1935-2009. Emerg Infect Dis 18:1248–1255. https://doi.org/10.3201/eid1808.120082
17. Belcher T, Preston A. 2015. *Bordetella pertussis* evolution in the (functional) genomics era. Pathog Dis 73:ftv064. https://doi.org/10.1093/femspd/ftv064
18. Mooi FR. 2010. *Bordetella pertussis* and vaccination: the persistence of a genetically monomorphic pathogen. Infect Genet Evol 10:36–49. https://doi.org/10.1016/j.meegid.2009.10.007
19. Bowden KE, Williams MM, Cassiday PK, Milton A, Pawloski L, Harrison M, Martin SW, Meyer S, Qin X, DeBolt C, Tasslimi A, Syed N, Sorrell R, Tran M, Hiatt B, Tondella ML. 2014. Molecular epidemiology of the pertussis epidemic in Washington State in 2012. J Clin Microbiol 52:3549–3557. https://doi.org/10.1128/JCM.01189-14
20. Advani A, Van der Heide HGJ, Hallander HO, Mooi FR. 2009. Analysis of Swedish *Bordetella pertussis* isolates with three typing methods: characterization of an epidemic lineage. J Microbiol Methods 78:297–301. https://doi.org/10.1016/j.mimet.2009.06.019
21. Cassiday PK, Skoff TH, Jawahir S, Tondella ML. 2016. Changes in predominance of pulsed-field gel electrophoresis profiles of *Bordetella pertussis* isolates, United States, 2000-2012. Emerg Infect Dis 22:442–448. https://doi.org/10.3201/eid2203.151136
22. Brinig MM, Cummings CA, Sanden GN, Stefanelli P, Lawrence A, Relman DA. 2006. Significant gene order and expression differences in *Bordetella pertussis* despite limited gene content variation. J Bacteriol 188:2375–2382. https://doi.org/10.1128/JB.188.7.2375-2382.2006
23. Stibitz S, Yang M-S. 1997. Genomic fluidity of *Bordetella pertussis* assessed by a new method for chromosomal mapping. J Bacteriol 179:5820–5826. https://doi.org/10.1128/jb.179.18.5820-5826.1997
24. Abrahams JS, Weigand MR, Ring N, MacArthur I, Etty J, Peng S, Williams MM, Bready B, Catalano AP, Davis JR, Kaiser MD, Oliver JS, Sage JM, Bagby S, Tondella ML, Gorringe AR, Preston A. 2022. Towards comprehensive understanding of bacterial genetic diversity: large-scale amplifications in *Bordetella pertussis* and *Mycobacterium tuberculosis*. Microb Genom 8:000761. https://doi.org/10.1099/mgen.0.000761
25. Weigand MR, Peng Y, Loparev V, Batra D, Bowden KE, Burroughs M, Cassiday PK, Davis JK, Johnson T, Juieng P, Knipe K, Mathis MH, Pruitt AM, Rowe L, Sheth M, Tondella ML, Williams MM. 2017. The history of *Bordetella pertussis* genome evolution includes structural rearrangement. J Bacteriol 199:e00806-16. https://doi.org/10.1128/JB.00806-16
26. Bowden KE, Weigand MR, Peng Y, Cassiday PK, Sammons S, Knipe K, Rowe LA, Loparev V, Sheth M, Weening K, Tondella ML, Williams MM. 2016. Genome structural diversity among 31 *Bordetella pertussis* isolates from two recent U.S. whooping cough statewide epidemics. mSphere 1:e00036-16. https://doi.org/10.1128/mSphere.00036-16
27. Weigand MR, Peng Y, Loparev V, Johnson T, Juieng P, Gairola S, Kumar R, Shaligram U, Gowrishankar R, Moura H, Rees J, Schieltz DM, Williamson Y, Woolfitt A, Barr J, Tondella ML, Williams MM. 2016. Complete genome sequences of four *Bordetella pertussis* vaccine reference strains from serum institute of India. Genome Announc 4:e01404-16. https://doi.org/10.1128/genomeA.01404-16
28. Weigand MR, Peng Y, Pouseele H, Kania D, Bowden KE, Williams MM, Tondella ML. 2021. Genomic surveillance and improved molecular typing of *Bordetella pertussis* using wgMLST. J Clin Microbiol 59:e02726-20. https://doi.org/10.1128/JCM.02726-20
29. Weigand MR, Williams MM, Peng Y, Kania D, Pawloski LC, Tondella ML, CDC Pertussis Working Group. 2019. Genomic survey of *Bordetella pertussis* diversity, United States, 2000-2013. Emerg Infect Dis 25:780–783. https://doi.org/10.3201/eid2504.180812
30. Kamachi K, Koide K, Otsuka N, Goto M, Kenri T. 2024. Strain identity of *Bordetella pertussis* isolates from household members based on whole-genome sequencing. Microbiol Immunol 68:111–114. https://doi.org/10.1111/1348-0421.13113
31. Beall B, Cassiday PK, Sanden GN. 1995. Analysis of *Bordetella pertussis* isolates from an epidemic by pulsed-field gel electrophoresis. J Clin Microbiol 33:3083–3086. https://doi.org/10.1128/jcm.33.12.3083-3086.1995
32. Fennelly NK, Sisti F, Higgins SC, Ross PJ, van der Heide H, Mooi FR, Boyd A, Mills KHG. 2008. *Bordetella pertussis* expresses a functional type III secretion system that subverts protective innate and adaptive immune responses. Infect Immun 76:1257–1266. https://doi.org/10.1128/IAI.00836-07
33. Gaillard ME, Bottero D, Castuma CE, Basile LA, Hozbor D. 2011. Laboratory adaptation of *Bordetella pertussis* is associated with the loss of type three secretion system functionality. Infect Immun 79:3677–3682. https://doi.org/10.1128/IAI.00136-11
34. Kamanova J. 2020. *Bordetella* type III secretion injectosome and effector proteins. Front Cell Infect Microbiol 10:466. https://doi.org/10.3389/fcimb.2020.00466
35. Armstrong GL, MacCannell DR, Taylor J, Carleton HA, Neuhaus EB, Bradbury RS, Posey JE, Gwinn M. 2019. Pathogen genomics in public health. N Engl J Med 381:2569–2580. https://doi.org/10.1056/NEJMsr1813907
36. Bouchez V, Guglielmini J, Dazas M, Landier A, Toubiana J, Guillot S, Criscuolo A, Brisse S. 2018. Genomic sequencing of *Bordetella pertussis* for epidemiology and global surveillance of whooping cough. Emerg Infect Dis 24:988–994. https://doi.org/10.3201/eid2406.171464
37. Payne M, Xu Z, Hu D, Kaur S, Octavia S, Sintchenko V, Lan R. 2023. Genomic epidemiology and multilevel genome typing of *Bordetella pertussis*. Emerg Microbes Infect 12:2239945. https://doi.org/10.1080/22221751.2023.2239945
38. Peng Y, Williams MM, Xiaoli L, Simon A, Fueston H, Tondella ML, Weigand MR. 2024. Strengthening *Bordetella pertussis* genomic surveillance by

Downloaded from https://journals.asm.org/journal/msystems on 15 October 2025 by 24.117.150.43.

direct sequencing of residual positive specimens. J Clin Microbiol 62:e0165323. https://doi.org/10.1128/jcm.01653-23
39. Didelot X, Walker AS, Peto TE, Crook DW, Wilson DJ. 2016. Within-host evolution of bacterial pathogens. Nat Rev Microbiol 14:150–162. https://doi.org/10.1038/nrmicro.2015.13
40. Hayden HS, Joshi S, Radey MC, Vo AT, Forsberg C, Morgan SJ, Waalkes A, Holmes EA, Klee SM, Emond MJ, Singh PK, Salipante SJ. 2022. Genome capture sequencing selectively enriches bacterial DNA and enables genome-wide measurement of intrastrain genetic diversity in human infections. mBio 13:e0142422. https://doi.org/10.1128/mbio.01424-22
41. Madi N, Chen D, Wolff R, Shapiro BJ, Garud NR. 2023. Community diversity is associated with intra-species genetic diversity and gene loss in the human gut microbiome. Elife 12:e78530. https://doi.org/10.7554/eLife.78530
42. Hardwick TH, Cassiday P, Weyant RS, Bisgard KM, Sanden GN. 2002. Changes in predominance and diversity of genomic subtypes of *Bordetella pertussis* isolated in the United States, 1935 to 1999. Emerg Infect Dis 8:44–49. https://doi.org/10.3201/eid0801.010021
43. Andrews S. 2010. FastQC: a quality control tool for high throughput sequence data. Cambridge, United Kingdom
44. Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J 17:10. https://doi.org/10.14806/ej.17.1.200
45. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. 2019. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol 37:540–546. https://doi.org/10.1038/s41587-019-0072-8
46. Krumsiek J, Arnold R, Rattei T. 2007. Gepard: a rapid and sensitive tool for creating dotplots on genome scale. Bioinformatics 23:1026–1028. https://doi.org/10.1093/bioinformatics/btm039
47. Seemann T. 2015. Snippy: fast bacterial variant calling from NGS reads. Web citation. https://github.com/tseemann/snippy.
48. Seeman T. 2018. snp-dists: pairwise SNP distances from FASTA sequence alignments. Web citation. https://github.com/tseemann/snp-dists.
49. Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:1–10. https://doi.org/10.1186/gb-2009-10-3-r25
50. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079. https://doi.org/10.1093/bioinformatics/btp352
51. Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. 2018. MUMmer4: a fast and versatile genome alignment system. PLoS Comput Biol 14:e1005944. https://doi.org/10.1371/journal.pcbi.1005944
52. Gates I, DuVall M, Ju H, Tondella ML, Pawloski L, Pertussis Working Group. 2017. Development of a qualitative assay for screening of *Bordetella pertussis* isolates for pertussis toxin production. PLoS One 12:e0175326. https://doi.org/10.1371/journal.pone.0175326
53. Weigand MR, Pawloski LC, Peng Y, Ju H, Burroughs M, Cassiday PK, Davis JK, DuVall M, Johnson T, Juieng P, Knipe K, Loparev VN, Mathis MH, Rowe LA, Sheth M, Williams MM, Tondella ML. 2018. Screening and genomic characterization of filamentous hemagglutinin-deficient *Bordetella pertussis*. Infect Immun 86:e00869-17. https://doi.org/10.1128/IAI.00869-17
54. Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34:3094–3100. https://doi.org/10.1093/bioinformatics/bty191
55. Li H. 2021. New strategies to improve minimap2 alignment accuracy. Bioinformatics 37:4572–4574. https://doi.org/10.1093/bioinformatics/btab705
56. Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, Schatz MC. 2018. Accurate detection of complex structural variations using single-molecule sequencing. Nat Methods 15:461–468. https://doi.org/10.1038/s41592-018-0001-7

Downloaded from https://journals.asm.org/journal/msystems on 15 October 2025 by 24.117.150.43.

Hannivoort et al. Fibrogenesis & Tissue Repair 2012, 5:1
http://www.fibrogenesis.com/content/5/1/1

Fibrogenesis & Tissue Repair

REVIEW **Open Access**

Genomics and proteomics in liver fibrosis and cirrhosis

Rebekka A Hannivoort[1,2], Virginia Hernandez-Gea[1] and Scott L Friedman[1*]

Abstract

Genomics and proteomics have become increasingly important in biomedical science in the past decade, as they provide an opportunity for hypothesis-free experiments that can yield major insights not previously foreseen when scientific and clinical questions are based only on hypothesis-driven approaches. Use of these tools, therefore, opens new avenues for uncovering physiological and pathological pathways. Liver fibrosis is a complex disease provoked by a range of chronic injuries to the liver, among which are viral hepatitis, (non-) alcoholic steatohepatitis and autoimmune disorders. Some chronic liver patients will never develop fibrosis or cirrhosis, whereas others rapidly progress towards cirrhosis in a few years. This variety can be caused by disease-related factors (for example, viral genotype) or host-factors (genetic/epigenetic). It is vital to establish accurate tools to identify those patients at highest risk for disease severity or progression in order to determine who are in need of immediate therapies. Moreover, there is an urgent imperative to identify non-invasive markers that can accurately distinguish mild and intermediate stages of fibrosis. Ideally, biomarkers can be used to predict disease progression and treatment response, but these studies will take many years due to the requirement for lengthy follow-up periods to assess outcomes. Current genomic and proteomic research provides many candidate biomarkers, but independent validation of these biomarkers is lacking, and reproducibility is still a key concern. Thus, great opportunities and challenges lie ahead in the field of genomics and proteomics, which, if successful, could transform the diagnosis and treatment of chronic fibrosing liver diseases.

Keywords: cirrhosis, genomics, liver fibrosis, mass spectrometry, microarray, proteomics

Introduction

Liver fibrosis results from a wound-healing response to chronic injury, which leads to excessive matrix, or scar deposition. This scar tissue can restrict blood flow due to contraction of the organ, leading to progressive liver damage and cirrhosis (the end stage of fibrosis), complicated by liver failure, portal hypertension and/or hepatocellular carcinoma [1]. Fibrosis is prominent in chronic liver diseases, including viral hepatitis, alcoholic and non-alcoholic steatohepatitis, toxic liver injury, autoimmune diseases and several genetic diseases. There have been two major priorities for therapy to reduce fibrosis: 1) to establish treatments for the diseases that lead to liver fibrosis; and, 2) to identify agents that directly slow or reverse fibrogenesis independent of the underlying disease.

A key discovery in understanding fibrosis has been the role of hepatic stellate cells (HSCs), vitamin A storing cells in the space of Disse, which, when activated, transform into myofibroblast-like cells, shedding their vitamin A content, and producing fibrogenic proteins, including collagens and tissue inhibitor of metalloproteinases-1 (TIMP-1) [2]. This review will focus on the contribution of high-throughput genomic and proteomic approaches to the study of fibrogenesis and fibrosis progression, concentrating on the most prevalent human chronic liver diseases and findings from animal models in liver tissue, isolated liver cells, cell lines and serum.

The role of genomics and proteomics in degenerative diseases and liver fibrosis

Genetic diseases can be classified as chromosomal abnormalities (for example, trisomy 21), Mendelian

---

* Correspondence: scott.friedman@mssm.edu
[1]Department of Medicine/Division of Liver Diseases, Mount Sinai School of Medicine, New York, NY 10029, USA
Full list of author information is available at the end of the article

© 2012 Hannivoort et al; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

disorders (single gene alterations with typical inheritance patterns, like autosomal dominant/recessive or X-linked), and complex diseases that are influenced by many genetic and environmental components. Degenerative diseases like liver fibrosis are complex illnesses [3]. The genetic contributions to these disorders are not attributable to a single gene alteration, but rather to a host of genetic susceptibilities defined by single nucleotide polymorphisms (SNPs) that predispose an individual to a disease. The susceptibility to an accumulation of environmental influences is either enhanced or reduced by genetic factors, thereby defining an individual's disease risk. Studies investigating these genetic traits are complicated, because there are many genes that influence the risk for complex diseases, yet the impact of each single genetic variant by itself is small. Therefore, large numbers of subjects are needed to provide sufficient statistical power to yield robust conclusions. Currently, there are almost 13 million SNPs catalogued in the NCBI human SNP database. Approaches to identify SNPs that are linked with a specific disease range from efforts to sequence specific disease-causing genes to genome scans requiring sequencing of large numbers of known SNPs that may or may not be associated with the disease.

Genomic and proteomic screening methods are often used to identify classes of genes that are differentially expressed in disease. These classes provide the investigator with potential pathways that could be involved in the regulation of this disease, thereby narrowing the focus of subsequent studies to uncover disease mechanisms and potential targets for therapy. By not limiting the study to pathways already associated with a disease, many new pathways not previously implicated are now under investigation [4].

This prospect is especially compelling in degenerative diseases like liver fibrosis, as these conditions often develop slowly, and therapies may not be necessary for all subclasses of patients [5]. Regardless of the etiology of liver fibrosis, some people progress rapidly towards cirrhosis, whereas others never develop fibrosis in the first place, or have slow progression of their fibrosis. It seems very unlikely that this phenomenon can be attributed solely to environmental influences. Identifying which patients are unlikely to ever progress to cirrhosis may prevent overtreatment of many patients.

Another use for these screening methods is to identify expression profiles, or patterns of expression of many genes, that in aggregate correspond with disease outcome or the response to therapy. By analyzing these different gene/proteome classes, investigators may ultimately predict an individual patient's response to therapy more accurately, or identify those who are at greatest risk of progression, thereby refining treatment decisions.

The need for non-invasive markers in the assessment of fibrosis progression

The need to assess hepatic fibrosis progression is becoming more important as the incidence of advanced liver disease continues to rise, especially due to viral hepatitis and fatty liver disease. Liver biopsies are currently the 'gold standard' for determining the presence and progression of liver fibrosis. They are performed percutaneously, via transjugular access, or during abdominal surgery. Liver biopsies have a low complication rate; however, significant hemorrhage may rarely occur, requiring hospitalization in approximately 2% of the patients; moreover, some patients have contraindications for liver biopsy [6-8]. Additionally, analysis of liver biopsy is often undermined by sampling error and inter-observer variation [[6,9-15]. Studying mRNA and protein disease profiles in liver biopsies is challenging, as tissue sample sizes are small, especially for protein analysis, and the biopsies contain a mix of many different cell types, including hepatocytes, hepatic stellate cells, Kupffer cells, endothelial cells, lymphocytes, red blood cells and bile duct epithelial cells. To study a single cell population from tissue sections requires laser-capture micro-dissection, which yields even less mRNA and protein than biopsy [15].

Close follow-up of liver disease by serial biopsy is not practical due to its invasive nature, fueling the need to find alternatives for liver biopsy. A less invasive and more accurate diagnostic tool would aid in diagnosis of fibrosis, follow-up of established therapies and evaluation of experimental therapies. Moreover, hepatocellular carcinoma (HCC) is one of the main complications of cirrhosis, and early detection markers that distinguish HCC from cirrhosis could significantly reduce morbidity and mortality. In the past ten years, genomic, transcriptomic and proteomic studies have provided candidate biomarkers for fibrosis assessment. Thus, in combination with standard laboratory serum analysis, these have led to novel fibrosis assessment tools using indirect and direct fibrosis markers, like FibroSpect (Prometheus Laboratories; San Diego, CA, US), FibroTest (BioPredictive;Paris, FR) and Fibroscan (Echosens; Paris, FR), ActiTest (BioPredictive;Paris, FR) and APRI (public domain) [16-20]. For a more exhaustive overview of currently used serum biomarkers, see a recent review by Smith *et al.* [21].

These fibrosis assessment tests have shown promising results, but a number of pitfalls are encountered, including limited availability, inter-laboratory variation, cost, insufficient resolution to differentiate between intermediate stages of fibrosis and false positive results due to other conditions, such as increased venous pressure or steatosis [22-29].

Because current methods for non-invasive assessment of liver fibrosis lack resolution, particularly in the intermediate fibrosis stages, more research is being directed towards genomics and proteomics in the search of biomarkers that provide a close association with fibrosis stage [30-35].

Technical aspects of genomics and proteomics in chronic liver disease

Genomics and transcriptomics

Genomics comprises the study of the genetic information (DNA, or RNA in certain viruses) of an organism. Transcriptomics aims to elucidate the transcripts of the genome or gene expression levels on the RNA level under varying conditions as changes in mRNA expression do not necessarily correspond with changes in protein expression, since alternative splicing, protein production, degradation and post-translational modifications influence protein stability and function [36,37].

The cDNA microarray is the most frequently used high-throughput screening method for gene expression profiling. These 'lab-on-a-chip' assays contain specified DNA oligonucleotide probes in spots on a chip that hybridize with oligonucleotides from tissue or cells providing a quantitative assessment of the transcribed genes [42-44]. Thousands of genes can be analyzed simultaneously with cDNA microarray and variants on this technique include specialized arrays that can detect SNPs and alternative splice variants (exon junction arrays) [38-41].

Serial analysis of gene expression (SAGE) analyzes the transcriptome by cloning strings of short cDNA fragments into bacteria, sequencing the cDNA fragments and then counting the number of cDNA fragments. This technique can yield information on both gene expression and alternative splicing and can help identify previously unknown genes [42,43].

Target genes that are revealed by the genomics/transcriptomics screens are generally validated by reverse transcription quantitative polymerase chain reaction (RT-qPCR), which is more quantitative and more reproducible than microarray, requires less mRNA and is more sensitive for genes that are expressed at low levels.

Proteomics

Proteomics is the study of the proteome in a cell compartment, tissue or organism comprising all proteins that are encoded for in the genome. Mass spectrometry (MS) is the key technique in proteomics. Many variants of mass spectrometers are currently in use, but they all rely on the same concept: to determine the accurate mass of a protein by measuring the mass-to-charge (m/z) ratio. In short, mass spectrometers consist of an ion source, a mass analyzer and a mass detector. After ionization, proteins and peptides travel through the mass analyzer, which evaluates their ratio of charge (z) versus mass (m). The mass detector counts the number of molecules per charge-to-mass ratio, and can provide the user with an output as a mass spectrum, with the m/z ratio on the x-axis and the molecule count per m/z on the y-axis [44].

MS is preceded by a separation step of proteins, either through 2D-polyacrylamide gel electrophoresis (2-DE/2D-PAGE), liquid chromatography (LC) or gel electrophoresis (DIGE) [45-47]. Once the protein spots are excised from the gel and digested, they can be processed through one mass analyzer (MS) to identify peptide mass, or two mass analyzers (MS/MS) to determine amino acid sequence.

MS requires ionization of proteins and peptides and two soft-ionizing techniques have been developed: matrix-assisted laser desorption/ionization (MALDI) and Electrospray ionization (ESI). Both techniques charge molecules that are subsequently analyzed by one or two mass analyzers (MS or versus tandem MS/MS). Time-of-flight (TOF) mass analyzers deduct peptide mass by measuring the time it takes for a charged peptide to travel through a vacuum tube in an electric field and can also be combined with one (Q-TOF) or two quadrupoles (Q-Q-TOF), which uses an oscillating electric field to selectively allow for molecules with a specific range of m/z ratios to proceed without collision [48].

When extremely high resolution is required, Fourier transform-ion cyclotron resonance (FT-ICR) MS or orbitrap can be used. The charged molecules are injected into a Penning ion trap where detectors measure the signal of molecules that pass. The m/z ratio determines the frequency with which the ion passes over the detector [49-51].

Surface-enhanced laser desorption/ionization (SELDI) analyzes protein mixtures that selectively bind to a biochip with the characteristics of choice. It needs only a small amount of crude sample (for example, serum or small liver biopsies), and is suitable for profiling multiple low-molecular-weight proteins. SELDI does not provide direct identification of the proteins. However, the peak profiles generated by SELDI can be useful by themselves in diagnostics, or in predicting prognosis and/or treatment response. Since manual labor is minimal, SELDI is by itself a suitable technique for high-throughput screening, especially for serum biomarkers; however, the cost is still too high for application on a large clinical scale [52-54]. For an excellent overview of the different types of mass spectrometers and their individual strengths and weaknesses, see a review by Domon and Aebersold [55].

A different type of high-throughput screening of proteins is the protein microarray. It functions much like a cDNA microarray, but probes are constructed out of proteins, antibodies or DNA constructs containing protein binding sequences that can capture proteins on a chip [56].

Technical pitfalls

Technical variations are largely responsible for the low reproducibility between studies by different research groups. Differences in patient characteristics can significantly alter genomic and proteomic profiles. Proteome studies are especially difficult because proteins cannot be amplified, unlike DNA and mRNA. Therefore, scientists have to rely on fractionation, enrichment/depletion and solubilization protocols prior to MS to prevent overload with the most abundant proteins present in the serum urine

Overview of genomics and proteomics in chronic liver disease

Most chronic liver diseases lead to fibrosis and cirrhosis in a significant subset of patients. While many common pathways drive fibrogenesis in all these diseases, there are also disease-specific pathways that contribute to fibrosis. An increasing number of studies are comparing the transcriptome and proteome of patients with different types of chronic liver injury to unearth disease-specific abnormalities in gene or protein expression.

In the following section we focus on disease-specific genomic and proteomic studies in both animal models and human diseases.

In vivo *and* in vitro *animal models of fibrosis*

Not all human liver diseases can be reproduced accurately in animal models, and thus surrogate models are being developed to study mechanisms of chronic liver injury and fibrosis.

Rodent studies reviewed below have analyzed a variety of fibrosis progression and resolution models, and have evaluated liver tissue, cell isolates, serum and urine proteomes. Considerable additional effort will be required to systematically compare proteomes between different disease models, and to accomplish the translational step of comparing these proteomic changes to human liver diseases.

Liver tissue proteome analysis

A number of liver tissue proteome analyses have been reported in rodent models. Low *et al.* characterized the proteome of rat livers treated with thioacetamide (TAA). Using several time points, they proposed that TAA causes chronic liver injury leading to liver fibrosis through down-regulation of enzymes related to fatty acid β-oxidation, branched chain amino acids and methionine breakdown through depletion of succinyl-CoA and subsequent alterations in heme and iron metabolism [57]. Using carbon tetrachloride (CCl_4) as a model for toxic liver injury in mice, and Abcb4-knockout mice as a model for sclerosing cholangitis to study differentially expressed proteins in liver tissue, Henkel *et al.* combined DIGE with MALDI-TOF MS and peptide mass fingerprint database search. They identified 20 differentially expressed genes in the CCl_4 model and 8 genes in the Abcb4-knockout model compared to control mice [58]. A number of studies have characterized proteome changes after ethanol feeding in rats. Shepard *et al.* observed that many mitochondrial proteins are acetylated in livers of rats that are chronically ethanol-fed [59]. A study in ethanol-fed rats treated with S-adenosylmethionine (SAM) demonstrated that the proteins affected by ethanol and SAM treatments were chaperones, beta oxidation proteins, sulfur metabolism proteins and dehydrogenase enzymes involved in methionine, glycine and choline metabolism [60]. Kharbanda *et al.* treated ethanol-fed rats with betaine, which restores the metabolic ratio of liver S-adenosylmethionine to S-adenosylhomocysteine. Ethanol significantly reduced carbonic anhydrase-III protein levels, which can lead to decreased resistance against oxidative stress [61].

Rodent studies are also being used as models for fibrosis resolution after the chronic injury is removed. Liu *et al.* characterized liver tissue proteome changes during spontaneous recovery after TAA-induced micronodular cirrhosis in rats. Using 2-DE and MALDI-TOF, they identified the up-regulation of GST-P2 that peaked after two weeks of recovery [62].

Employing a cDNA microarray gene chip and 2D-DIGE with MALDI-TOF/MS-MS, Kirpich *et al.* found that glutathione S-transferases mu-1, pi-1 and selenium-binding protein 2 are decreased at both gene and protein levels in a mouse model of NAFLD [63].

Enoyl-coenzyme A hidratase, an enzyme that catalyzes the second step of mitochondrial fatty acid beta-oxidation, has been described by Zhang *et al.* to be down-regulated in rats with HFD-induced hepatic steatosis [64].

Hepatic stellate cell proteome analysis

Activation of HSCs is a key event in fibrogenesis and can be reproduced *in vitro* by prolonged culture on plastic or collagen. The changes in protein expression during activation can provide novel targets for antifibrotic therapy. However, differences between *in vitro* and *in vivo* HSC activation need to be taken into account. Kristensen *et al.* isolated hepatic stellate cells from rats and compared the proteome of quiescent cells with that of either culture-activated cells or *in vivo* activated cells from rats treated with CCl_4 for eight weeks. Following 2-DE, proteins were digested by trypsin, and ESI-MS/MS was used to sequence the proteins that were differentially expressed. They detected 16 proteins showing differences based on the model of activation [65].

A study by Kawada *et al.* explored the proteome of *in vivo* (thioacetamide) and *in vitro* activated rat HSCs, describing a protein that was induced with HSC activation: stellate-cell activation-associated protein (STAP), which may act as an anti-fibrotic scavenger of peroxides during liver injury [66].

Deng *et al.* employed human hepatic stellate cells (LX2) treated and untreated with taurine and performed a combination of two-dimensional gel electrophoresis and ultra-performance liquid chromatography-electrospray ionization-tandem mass spectrometry (UPLC-ESI-MS/MS). They postulate a beneficial role of taurine in hepatic fibrosis, as they were able to detect an increased rate of stellate cell apoptosis after treatment [67].

Serum and urinary proteome analysis

Animal models could provide general serum markers that correlate with fibrosis stage and have the advantage of easier access to liver tissue at set time points for histological comparison. In thioacetamide (TAA) and bile-duct ligation treated rats, Xu *et al.* indentified and sequenced a 3.5 kDa histidine-rich glycoprotein, which had > 92% specificity and > 97% sensitivity for identifying liver cirrhosis. Previous studies have shown a beneficial effect of bone marrow cell (BMC) transplantation in CCl_4-induced cirrhosis in mice, improving liver regeneration, function and fibrosis. Yokoyama *et al.* found six differentially expressed proteins in the serum by 2-DE, 48 hours after BMC transplantation, compared to non-transplanted mice. They suggested ApoA1 protein levels as a proteomic analysis of serum marker proteins in recipient mice with liver cirrhosis after bone marrow cell transplantation. marker for liver regeneration after BMC transplantation [68]. Serum taurocholic acid has also been recently proposed as a marker of early hepatic damage based on the Shimada *et al.* study. The identified elevated serum taurocholic acid concentration by SPE-MALDI-TOF MS in mice CCl_4 induced liver injury [69].

A recent publication looked at urinary samples to find biomarkers for CCl_4-induced liver fibrosis providing a potential alternative to serum samples for non-invasive diagnostics and follow-up in liver fibrosis (70).

Chronic hepatitis C viral infection (HCV)

In patients with HCV, early knowledge of which patients have rapidly progressing fibrosis could be very beneficial in the decision-making process of whether to treat patients aggressively with antiviral therapy. A recent review by Walters and Katze explored the relationship among the gene expression profile of HCV patients, viral clearance and treatment response, shedding light on potential virus-host interactions, which may emerge as future therapeutic targets [71].

Staging of fibrosis

The liver proteome of HCV-patients can provide HCV-specific and general fibrotic pathways that could lead to novel therapeutic targets, or predict the natural disease progression or therapeutic response in the individual patient. Recent findings point towards alterations in fatty acid oxidation, oxidative phosphorylation and structural proteins in advanced fibrosis. An overview of high-throughput studies analyzing pathways and proteins involved in fibrosis due to HCV is represented in Table 1.

Diamond *et al.* have described a pattern of 210 proteins that correlate with fibrosis stage in 1,641 HCV-infected patients. They were able to cluster the patients in fibrosis stages 3 to 4 versus stages 1 to 2 using these protein expression profiles. Functional analysis indicated reduced expression of genes involved in fatty acid oxidation and oxidative phosphorylation in advanced fibrosis [72]. Several cell structure-associated proteins were uncovered in the cirrhotic septa of seven HCV patients (METAVIR stage F4) when Mölleken *et al.* compared the proteomic profiles of microdissected cirrhotic septa versus parenchymal liver cells using 2-DE and LC with ESI-MS/MS, in the search for new serum biomarkers for fibrosis. MFAP-4 could predict the presence of fibrosis with high accuracy in both HCV- and alcohol-induced fibrosis, but diagnostic accuracy decreased significantly for the discrimination between fibrosis stages 2 to 4 [73].

Regarding serum markers, none of these serum markers can adequately distinguish intermediate fibrosis stages, even though the progression from mild to intermediate fibrosis is an important predictor of further fibrosis progression. White *et al.* reported three up-regulated proteins (α2-macroglobulin, haptoglobin, albumin) and four down-regulated proteins (complement C-4, serum retinol binding protein, apolipoprotein A1, and two isoforms of apolipoprotein A4) in the serum of HCV patients with advanced fibrosis (METAVIR stage F3/4, n = 23) versus no or mild fibrosis (METAVIR stage F0/1, n = 21) [74]. Three of these proteins are already being used in the FibroTest [75]. This study was limited by a small sample size, and the four new potential biomarkers could not be validated in a larger patient group. Gangadharan *et al.* found a decrease in several proteins, including apolipoprotein L1, prealbumin, albumin, haptoglobin and complement components, whereas other proteins, such as CD5 antigen-like (CD5L), 2-macroglobulin and immunoglobulin components, were induced in fibrosis when they studied the serum proteome of HCV patients with varying levels of fibrosis and healthy controls, using 2-DE and LC with Q-TOF MS/MS [76].

Table 1 High-throughput studies in HCV patients analyzing differential gene expression based on fibrosis stage

1st author	Year	# Patients	# Controls	Specimen	Methods	# Differentially expressed genes	Main pathways involved	Most differentially regulated genes	⇑/⇓	Reference
Huang	2006	121	312	Blood	SNP microarray	1,609	N/A	DEAD-box popypeptide 5 (DDX5) SNP Carnitine palitoyltransferase 1A (CPT1A) SNP	⇑ ⇑	81
Diamond	2007	22	4	Liver tissue	LC-ESI-FTICR MS/MS	210	Amino acid, carbohydrate and lipid metabolism Oxidative phosphorylation Oxidative stress	N/A	N/A N/A N/A	72
White	2007	23	21	Serum	2-DE/LC-Q-TOF MS/MS	7	N/A	Alpha-2-macroblobulin Haptoglobin Fragment of albumin Serum retinol binding protein (SRBP) Complement 4 Apolipoprotein A-I (Apo A1) Apolipoprotein A-IV (Apo A4)	⇑ ⇑ ⇑ ⇓ ⇓ ⇓ ⇓	74
Gangadharan	2007	11	4	Serum	2-DE-Q-TOF/CapLC MS/MS	83	Plasmin-associated Hepatic synthetic function HGF-releated Lipid metabolism Immune system related	Alpha-2-macroblobulin Inter-alpha-trypsin inhibitor heavy chain 4 Albumin Transthyretin Complement C3 and C4 Factor H-related protein 1 Alpha-1 antichymotrypsin Haptoglobin Apolipoprotein L-I Beta 2 glycoprotein CD5 antigen like IgA1+IgG2 heavy chain and Ig light chain regions	⇑ ⇓ ⇓ ⇓ ⇓ ⇓ ⇓ ⇓ ⇓ ⇓ ⇑ ⇓ ⇓ ⇑ ⇑	76
Mölleken	2009	7	7	microdissections	2-DE/LC-ESI-Ion trap MS/MS	35	Cell structure	Tropomyosin alpha Tropomyosin beta Microfibtil-associated glycoprotein 4 Calpoin 1 Transgelin Tropomyosin alpha 4 Actin alpha 1 Vimentin	⇑ ⇑ ⇑ ⇑ ⇑ ⇑ ⇑ ⇑	73

A novel set of four cDNA peripheral blood mononuclear cell markers (a2-macroglobulin, haptoglobin, mitogen-activated protein kinase kinase 3, and alanine aminopeptidase N) accurately predicted the stage of liver fibrosis in HIV/HCV co-infected patients with an area under the receiver operating curve (AUC ROC) of 0.852 [77].

L Yang *et al*. recently analyzed 24 liver samples from patients with chronic HCV by DIGE and Isobaric tags for relative and absolute quantitation (iTRAQ) and identified 2 proteins as biomarker candidates for predicting hepatic fibrosis: complement C4-A and inter-alpha-trypsin inhibitor heavy chain H4 [78,79].

Novel biomarker candidates to predict hepatic fibrosis in hepatitis c identified by serum proteomics

Fibrosis progression has also been related to galectin-3-binding protein (G3BP). Cheung *et al*. found increased levels in patients with hepatitis C-related cirrhosis when compared with mild and moderate fibrosis in 76 serum and 20 tissue human samples [80].

Prediction of fibrosis progression

Prediction of fibrosis progression is crucial for adequate patient selection in clinical trials, and, thus, Huang *et al.* studied the genome of 1,468 patients with chronic hepatitis C infection and known fibrosis stage. They developed a Cirrhosis Risk Score (CRS) consisting of SNPs in seven genes, which could predict fibrosis progression significantly better than clinical parameters. Four SNPs were in known genes: antizyme-inhibitor-1, toll-like receptor 4, transient receptor potential cation channel, subfamily M, member 5 and aquaporin 2 [81]. This initial study was independently validated two additional studies [82,83], and is especially exciting, as it is the first evidence that host genetics significantly influence fibrosis progression and, thereby, provides a means for selecting patients that will benefit most from aggressive antiviral or antifibrotic treatment.

At the proteomic level, when plasma from cirrhotic HCV patients is analyzed by SDS-PAGE and MS, both chains of the complement C3 and C4 show a significant decrease in comparison with samples from healthy patients [84,85].

Prediction of treatment response

Transcriptomic and proteomic studies point to a crucial role of interferon signaling pathways in the prediction of treatment response, where high pre-treatment interferon signaling is associated with poor virological response. Recent genomic studies have identified SNPs in several interferon-related genes, especially interleukin 28b, that are powerful predictors of virologic response indicating that genomics can be a powerful tool in the prediction of treatment response.

A recent publication by Asselah *et al.* demonstrated that liver gene expression can predict outcome of patients with chronic hepatitis C infection after treatment with pegylated interferon and ribavirin. They studied cDNA of non-responders (NR), and SVR using RT-PCR for 59 genes. Three genes were significantly induced in NRs versus SVRs: interferon alpha-inducible protein 3, interferon alpha-inducible protein 27 and interferon alpha-inducible protein 2. Supervised class prediction analysis identified a two-gene (interferon-alpha-inducible protein 27 and CXC chemokine ligand 9) signature, with a predictive accuracy of 100% in NRs and of 70% in SVRs. Although no high-throughput analysis was performed, this paper provides a basis for further exploring the use of genomic and proteomic information to predict individual treatment response, and may be useful for selecting those patients who will benefit from treatment [86].

Chen *et al*. analyzed liver tissue using cDNA microarray to identify a gene signature that can predict treatment outcome. A key finding was that interferon stimulated gene 15 (ISG15) protein induction was more pronounced in hepatocytes in non-responders, but in Kupffer cells in responders [75]. In a liver tissue microarray study by Hayashida *et al*., a different set of genes was found to predict treatment response in patients treated with interferon alone, or interferon/ribavirin co-treatment. In the interferon monotherapy patients, main predictors were genes involved in the interferon signaling pathway, lipid metabolism, complement and oxidation/reduction. In the combined treatment group, genes, including *cylophylin A*, and multidrug resistance protein-hepatic gene expression during treatment with peginterferon and ribavirin emerged, identifying molecular pathways for treatment response. 1 (P-gp) [87]. Feld *et al.* discovered that low baseline liver tissue cDNA levels of interferon-stimulated genes (ISGs) predict rapid treatment response with quick induction of ISG levels, whereas high baseline ISG levels do not increase much further and predict slow treatment response, together with up-regulation of interferon inhibitory pathways [88].

As described above, recent genome-wide association studies (GWAS) have identified SNPs in the interleukin 28b (IL28B, coding for interferon-γ-3) genomic region that are associated with viral response to the current treatment of HCV, pegylated-interferon-alpha (PEG-IFNα) and ribavirin. GWAS by Suppiah *et al.* led to the discovery of a SNP near interleukin 28b (IL28B) that is associated with sustained viral response (SVR) in genotype 1 HCV patients treated with PEG-IFNα and ribavirin, with an odds ratio (OR) for the G-allele of rs8099817 of 1.64 for heterozygous carriers and 2.39 for homozygous carriers compared to non-carriers [89]. Other SNPs that were associated with SVR included Caspase 1, a known activator of interleukin 1, and interleukin 21 receptor. A simultaneous report was published by Tanaka *et al.*, identifying SNPs in IL28B in a Japanese genotype 1 HCV-patient cohort as a predictor of virologic response to PEG-IFNα and ribavirin. Initial GWAS identified rs2980275 and rs8099917 SNPs as predictors of non-response and SVR. They continued to seek SNPs in the IL28A/B region and uncovered rs8105790, rs11881222, rs8103142, rs28416813, rs4803219, rs8099917 and rs7248668 as associated with virologic response [90]. Identification of those HCV-patients who are likely to respond to therapy could lead to a shortened duration of therapy in patients with a favorable SNP genotype. Although these studies did not examine fibrogenic end-points, they are a major breakthrough in HCV treatment response research, and are excellent examples of how GWAS can identify genes and SNPs that are related to therapeutic responses.

Non-responders generally have high expression of interferon-stimulated genes (ISGs) pre-treatment that cannot be further induced by PEG-IFNα, whereas RVR patients have strong induction of ISGs on treatment. Refractory high endogenous interferon signaling in non-responders may interfere with PEG-IFNα treatment, especially in genotype 1 and 4 patients. Whether inhibition of endogenous interferon signaling could improve therapeutic response is subject to further study [91,92].

Although further validation is needed, Devitt E *et al.* using SELDI-TOF MS analyzed serum from 25 HCV infected patients during the initial 24 h of treatment with pegylated interferon α-2b and ribavirin and identified 16 differential peaks able to distinguish responders from nonresponders [93].

Chronic hepatitis B viral infection (HBV)

Approximately 400 million patients worldwide are chronically infected with hepatitis B virus (CHB). Hepatitis B virus infection causes chronic hepatitis in about 10% of infected patients, and increases the risk of developing HCC about 100-fold versus in the uninfected population. An excellent review on the role of genomics, transcriptomics and proteomics in chronic HBV-associated liver disease, recently published by Sun *et al.*, described determinants of susceptibility to persistent HBV infection, disease progression and HCC development; their review also describes the use of these technologies to understand the pathogenic mechanism of HBV and to identify biomarkers that could aid in early detection of HCC [94].

Recent high-throughput proteomic studies have identified protein patterns that correlate with fibrosis stage in HBV patients. Zhang *et al.* described a novel approach to evaluate the protein expression profile of plasma membrane and analyzed human liver samples from HBV patients in different fibrosis stages. Using 2-DE, IPGphor isoelectronic focusing system and Bio-Rad Protein II electrophoresis to the investigators identified positive correlation between fibrosis grade and annexin A2 levels [95]. Lu *et al.* analyzed plasma of 7 healthy controls and 27 HBV patients with different stages of fibrosis by 2-DIGE and identified the up-regulation of peroxiredoxin 2 as a potential biomarker of HBV related liver fibrosis [96]. Although not liver-specific, plasma gelsolin protein has been pointed out as a potential predictor of HBV progression. C Marrocco *et al.* found by 2-DE a repressed expression of the protein in plasma of eight cirrhotic HBV patients, compared to eight chronic HBV infected patients [97]. Mohamadkhani *et al.* performed a 2DE and mass spectrometry analysis in patients with either chronic hepatitis B or HBV-related cirrhosis and healthy controls. Their data suggest that progression of HBV-related liver injury is associated with a down-regulation of apolipoprotein A1 and an increase in myeloperoxidase levels [98]. However, these results still need to be validated, and their ability to distinguish intermediate fibrosis stages is not clear.

Poon *et al.* using SELDI ProteinChip arrays and the Significance Analysis of Microarray (SAM) algorithm generated a model able to predict fibrosis (Ishak score ≥ 3) and cirrhosis (Ishak score ≥ 5) with an accuracy of > 90% [99]. Liver histology (METAVIR score) of HBV-patients was compared with three sets of serum markers by Myers *et al.*: aminotransferases (ALT, AST), Fibrotest (α_2-macroglobulin, apolipoprotein A1, haptoglobin, gamma-glutamyl-transpeptidase (GGT) and total bilirubin) and Actitest (ALT). Aminotransferases and Actitest were good predictors of activity (stage A2-F3) and Fibrotest accurately predicted stage F2-F4 fibrosis, especially in the ranges of ≤ 0.20 or > 0.80. Limiting liver biopsy to patients with intermediate Fibrotest scores could prevent biopsies in about half of the patients, without affecting accuracy [100].

Wang *et al.* performed an *in vitro* study to investigate proteome differences between stably HBV transfected HepG2.2.15 versus wild type HepG2 cells by 2-DE followed by LC-ESI-MS/MS. Furthermore; they compared HepG2.2.15 cells with and without interferon-alpha (IFN-α) treatment. Proteins that were differentially expressed in HepG2.2.15 versus HepG2 cells could be classified as being involved in, for example, cytoskeletal matrix, heat shock stress, signal transduction and protease/proteasome components. IFN-α treatment induced expression of interferon-stimulated gene 15 (ISG15) and prohibitin, among others, providing potential pathways that could lead to tumor suppression and defense against viral infection. Although these results have not been validated *in vivo*, they nevertheless establish that proteomics and cell culture can be a powerful combination to uncover differences in the proteome, and can be useful for evaluating potential treatments [101].

Alcoholic liver disease (ALD)

Several genomic and proteomic studies have been published on patients with alcoholic liver disease. However, the majority of studies focus on either diagnostics or genes/proteins that differentiate alcoholic hepatitis from alcoholic steatosis, not in predicting progression towards fibrosis and cirrhosis. Nevertheless, alcoholic steatohepatitis is clearly a precursor for developing fibrosis. For example, 586 genes were differentially expressed between alcoholic hepatitis and alcoholic steatosis, and 211 genes were differentially expressed between alcoholic hepatitis and non-disease controls when Seth *et al.* compared liver tissue RNA from eight patients with alcoholic hepatitis, nine with alcoholic steatosis, and seven non-disease controls using microarray analysis. Induced genes included pathways controlling cell adhesion, immune response, oncogenesis, signal transduction and embryogenesis. The 111 down-regulated genes included pathways for protein synthesis, cell growth and maintenance, transcription, signal transduction and transport [102].

Non-alcoholic fatty liver disease (NAFLD) and non-alcoholic steatohepatitis (NASH)

The prevalence of NAFLD and NASH is on the rise, which makes these increasingly important causes of liver fibrosis, especially in countries with a Western diet containing a high fat and/or carbohydrate content. For NAFLD and NASH there is also a lack of genomic/proteomic studies specifically examining fibrogenesis and fibrosis progression, much like in alcoholic liver disease. Most genomic/proteomic studies focus on distinguishing NASH from NAFLD, and not on progression to liver fibrosis and cirrhosis. The ability to distinguish NASH from NAFLD is very relevant, however, as NASH clearly predisposes to fibrosis. Therefore, we reviewed recent publications on this topic. However, studies directed at identifying proteins related to the speed of fibrosis progression in NASH could further select patients who could benefit from treatment, improve understanding of fibrosis pathogenesis in fatty liver, and identify targets for treatment of NASH-related fibrosis.

Staging of steatosis

NASH is distinguished from NAFLD by an inflammatory component. Therefore, not surprisingly, genes that are up-regulated in NASH versus NAFLD have strong relationships with inflammatory and immune pathways. Complementary DNA microarray analysis on nine liver tissue samples from steatotic livers and nine normal livers unearthed 110 differentially expressed genes, many related to mitochondrial respiratory and mitochondrial metabolic pathways. Genes belonging to the interleukin-1 receptor family and transforming growth factor-beta were induced in steatotic livers, pointing to a deranged inflammatory pathway, even in early development of steatosis. Thirty-four genes with significant differential expression were identified in patients with NASH, when compared with non-obese controls in a study by Younossi *et al.*, using cDNA microarray analysis on liver biopsies taken during bariatric surgery to find differential expression patterns between NASH (n = 29), NAFLD (n = 12), obese non-steatotic (n = 7), and non-obese controls (n = 6) [103]. Nineteen of these genes showed no significant expression differences in obese versus non-obese controls, suggesting a stronger association of these genes to NASH. These genes were related to lipid metabolism, extracellular matrix remodeling, liver regeneration, apoptosis and detoxification. In another study, this group used parallel liver tissue gene expression and serum proteome analysis in a group of 98 morbidly obese patients. Twenty-seven patients had developed NASH, 52 patients had steatosis with nonspecific inflammation, 12 patients had steatosis alone and 7 patients had normal histology on biopsy. Using SELDI-TOF analysis on serum samples, this group found several protein peaks associated with different stages of NAFLD/NASH. However, except for fibrinogen γ, the corresponding proteins were not identified. On the mRNA level, insulin-like growth factor binding protein 1 (IGFBP1) and fatty acid CoA ligase 4 (FACL4) were differentially expressed between NASH and steatotic patients.

A recent study of the serum proteome by Bell *et al.* led to the discovery of 55 proteins that changed significantly between simple steatosis and NASH with stage F3/F4 fibrosis (advanced bridging fibrosis/cirrhosis), and 15 proteins that changed between early NASH and NASH with stage F3/F4 fibrosis. Most of these proteins are involved in the immune response, coagulation, cellular and extracellular matrix and blood carrier proteins [104]. These proteins shed light on the pathogenesis of NAFLD/NASH and could be explored as candidates for biomarkers of advancing disease. Ulukaya *et al.* were able to use serum MALDI-TOF MS peaks to accurately predict NAFLD versus controls, but could not distinguish between simple steatosis and NASH. They included 80 patients with NAFLD, and 19 controls [105]. Histologic classification of the NAFLD patients showed definite NASH in 48 patients, borderline NASH in 22 patients and 10 patients with simple steatosis. Using reverse phase protein microarray technology, Calvert *et al.* revealed deranged insulin signaling in NAFLD patients when they investigated omental adipose tissue from 99 obese patients. Liver biopsies were classified as follows: (1) no fatty liver disease present, (2) simple steatosis, (3) steatosis with nonspecific inflammation, or (4) NASH. Categories 2 to 4 were clustered as NAFLD. Protein kinase A (PKA) and AKT/mTOR deregulation and caspase 9 activation were good predictors of NASH on liver histology compared to non-progressive forms of NAFLD [106]. For more information on genomic and proteomic analysis of NAFLD and NASH, see a recent review by Wilfred de Alwis and Day [107].

Fibrosis progression in steatotic patients

Very few publications report changes in gene expression that correlate with fibrosis stage in NAFLD/NASH patients. So far, only apolipoproteins, CD5 antigen-like (CD5L) and lumican have been identified. Gray *et al.* applied 2-DE and MALDI-TOF to analyze differences in serum proteome profile between patients with pre-cirrhotic NAFLD (n = 5), cirrhotic NAFLD (n = 5) and cirrhotic NAFLD with HCC (n = 5). Four out of five spots that identified at least one of the three groups were identified as apolipoprotein isoforms. The fifth spot, identified as CD5L, was induced in cirrhotic patients and even more in patients with HCC. ELISA analysis of CD5L in a larger patient group showed a poor AUROC score for distinguishing patients with HCC from those without HCC. However, as a cirrhosis biomarker, ROC analysis showed an accuracy of 72%. CD5L could potentially be used in combination with other biomarkers for cirrhosis [108]. Charlton *et al.* identified lumican, a 40 kDa keratin sulfate proteoglycan, as significantly increased in liver tissue from patients with mild NASH (NASH with F0/F1 fibrosis) versus simple steatosis, and in patients with progressive NASH (NASH with F2-F4 fibrosis) versus mild NASH in obese patients, when analyzed with LC-Q-TOF MS/MS. FABP-1 was increased in simple steatosis patients compared to non-steatotic patients, but was paradoxically decreased in NASH patients, which suggests an impaired detergent effect against free fatty acids in progressive NAFLD patients [109].

Autoimmune liver diseases

Primary biliary cirrhosis (PBC), as well as related autoimmune disorders, is more prevalent in certain families and thus seems to have a complex genetic component, with many factors probably adding to the risk of developing PBC. Some of these genes associated with PBC are major histocompatibility complex genes and common autoimmune genes [110]. Shackel *et al.* discovered that PBC liver tissue had increased expression of Th1 and Th2 type proteins as well as connective tissue growth factor and TGF-β3 when they analyzed their cDNA microarray from liver tissue from patients with PBC (n = 4), PSC (n = 4) and non-disease controls (n = 8). Moreover, Wnt and Notch signaling pathways were induced. PSC patient samples had increased expression of apoptosis associated molecules [111]. Serum samples from 44 PBC patients and 75 controls, analyzed using WCX magnetic beads and MALDI-TOF-MS, provided 69 potential protein biomarkers. Li *et al.* used these to construct a diagnostic model for PBC using the *m/z* peaks of protein biomarkers 3445, 4260, 8133, and 16290, which showed a sensitivity of 92.9% and a specificity of 82.4% [112].

Tahiri *et al.* noticed an induction of six known potential plasma membrane expressions: liver arginase, cytokeratins 8/18, heatshock proteins (HSP) 60/70/90, and valocin-containing protein (VCP) in autoimmune hepatitis when they investigated serum samples from 65 autoimmune hepatitis type 1 patients and 90 controls consisting of healthy blood donors (n = 40) and patients with systemic diseases (n = 20) or other liver diseases (n = 30). They proposed that these proteins could be targets for auto-antibodies in autoimmune hepatitis [113]. Using protein microarray, Song *et al.* analyzed serum samples from patients with autoimmune hepatitis (n = 44), PBC (n = 50), HCV (n = 41), HBV (n = 43), systemic lupus erythematosus (SLE) (n = 11), primary Sjögren syndrome (n = 11), rheumatoid arthritis (n = 2) and 50 healthy controls. Initially they found 11 differentially expressed genes between autoimmune hepatitis and other samples. They then produced an autoimmune hepatitis-specific protein chip, and found three new antigens, ribosomal protein S20 (RPS20), Alba-like, and dUTPase, that could be used as highly specific biomarkers for autoimmune hepatitis, and validated this with ELISA [114]. Bowlus *et al.* have recently analyzed by *in situ* MALDIMS (matrix-assisted laser desorption/ionization mass spectrometry) in PBC (n = 29), PSC (n = 11), AIH (n = 7) and healthy controls (n = 10) and found a promising pattern of protein expression in bile ducts, inflammatory infiltrates and hepatocytes that may represent a feasible way to identify novel targets in these diseases [115]. Song *et al.* created a protein microarray containing 5,011 recombinant human liver proteins and were able to identify three new antigens, RPS20, Alba-like and dUTPase, as highly autoimmune hepatitis-specific biomarkers [114].

Although these results are promising, the number of high-throughput genomic and proteomic studies in autoimmune liver diseases is still very limited.

Hepatocellular carcinoma

While beyond the scope of this review, there have been significant advances in early diagnosis, treatment and pathway analysis of HCC and hepatocarcinogenesis through genome, transcriptome and proteome high-throughput analysis, especially in the setting of liver cirrhosis. For more information on this topic we refer to recent reviews [116-118].

General conclusions and future directions

In the last decade, genomic and proteomic hypothesis-free studies have become increasingly common. Since liver fibrosis often takes many years to develop, and not every patient will become symptomatic, a method to select patients who will benefit from aggressive treatment in an early phase would be very beneficial. Therefore, there is a demand for biomarkers that can provide a prognostic indication for fibrosis, either in the natural course of chronic liver disease, or as a predictor of treatment response. Liver biopsy is currently the gold standard for follow-up in chronic liver disease, but it can have serious complications and suffers from sampling error and inter-observer variation. This drives the search for non-invasive direct or indirect biomarkers that correlate accurately with fibrosis stage.

Prospective predictive proteomic and genomic studies are difficult in liver disease, as follow-up is long, and biological samples can degrade over time. Thus far, many tests based on serum biomarkers correlate fairly well with the fibrosis stage when advanced fibrosis (METAVIR F3-4) is compared to mild or no fibrosis (F0 to 2), but lack resolution in the intermediate stages. Especially, these intermediate stages (F1 to 2) could be helpful in the early detection of progressive fibrosis and guide the hepatologist in treatment decisions. So far, genomic and proteomic studies for fibrosis staging have been mostly limited to viral hepatitis, with other diseases like alcoholic liver disease and NASH yet to be explored. Studying the fibrotic genome/proteome in these diseases will shed light on common and distinct pathological pathways leading to liver fibrosis.

Genomic and proteomic studies that have identified genes and proteins that correlate with fibrosis levels have yielded heterogeneous findings, most likely due to variations in sample preparations and patient populations/characteristics. The lack of standardized sample preparation protocols hinders reproducibility of protein spectra. A few gene and protein groups seem to correlate with fibrosis more consistently, mostly belonging to glutathione, oxidative stress, inflammatory and immune response pathways. For most of the identified proteins, adequate follow-up has been lacking so far, leaving scientists with lots of potential biomarker candidates still in need of validation.

Future high-throughput research will continue uncovering genes and proteins that can elucidate common and specific-disease pathways that lead to liver fibrosis. Standardization of sample preparation protocols should increase reproducibility of these studies, and will make it easier to select candidate genes and proteins that are suitable for validation as biomarkers for either fibrosis staging, or as predictive biomarkers for disease outcome and/or treatment response.

Abbrevations

2-DE/2D-PAGE: 2D-polyacrylamide gel electrophoresis; AIH: autoimmune hepatitis; ALD: alcoholic liver disease; BMC: bone marrow cell; CCl_4: carbon tetrachloride; CRS: Cirrhosis Risk Score; DIGE: gel electrophoresis; ESI: electrospray ionization; FACL4: fatty acid CoA ligase 4; FT-ICR: Fourier transform-ion cyclotron resonance; GWAS: genome-wide association studies; HCC: hepatocellular carcinoma; HCV: chronic hepatitis c viral infection; HSCs: hepatic stellate cells; IGFBP1: insulin-like growth factor binding protein 1; IMAC: immobilized metal affinity capture; ISGs: interferon-stimulated genes; iTRAQ: isobaric tags for relative and absolute quantitation; LC: liquid chromatography; MALDI: matrix-assisted laser desorption/ionization; MALDIMS: matrix-assisted laser desorption/ionization mass spectrometry; MS: mass spectrometry; m/z: mass-to-charge; NAFLD: non-alcoholic fatty liver disease; NASH: non-alcoholic steatohepatitis; OR: odds ratio; PBC: primary billiary cirrhosis; PCR: polymerase chain reaction; PKA: protein kinase A; Q-Q-TOF: second quadruple time-of-flight; Q-TOF: quadruple time-of-flight; RT-qPCR: reverse transcription quantitative PCR; SAGE: serial analysis of gene expression; SAM: significance analysis of microarray; SAX: strong anion exchange; SELDI: surface-enhanced laser desorption/ionization; SNPs: single nucleotide polymorphisms; STAP: stellate-cell activation-associated protein; SVR: sustained viral response; TAA: thioacetamide; TIMP-1: tissue inhibitor of metalloproteinases-1; TOF: time-of-flight; UPLC-ESI: ultra-performance liquid chromatography-electrospray ionization-tandem; WCX: weak cation exchange

Acknowledgements

RH was funded by the Graduate School for Drug Exploration (GUIDE), and the Stichting Nicolaas Muleriusfonds, Groningen, the Netherlands. SF and VH-G are supported by NIH Grants DK37340, DK56621 and AA017067.

Author details

[1]Department of Medicine/Division of Liver Diseases, Mount Sinai School of Medicine, New York, NY 10029, USA. [2]Department of Gastroenterology and Hepatology, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands.

Authors' contributions

RH participated in the design of the study, carried out the bibliographic search, drafted the manuscript and created the tables. VHG participated in the design of the study, the bibliographic search and the manuscript drafting. SF conceived of the study, participated in its design and coordination, and helped to draft the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Received: 28 October 2011 Accepted: 3 January 2012
Published: 3 January 2012

References

1. Bataller R, Brenner DA: **Liver fibrosis.** *J Clin Invest* 2005, **115**:209-18.
2. Friedman SL: **Hepatic stellate cells: protean, multifunctional, and enigmatic cells of the liver.** *Physiol Rev* 2008, **88**:125-72.
3. Juran BD, Lazaridis KN: **Applying genomics to the study of complex disease.** *Semin Liver Dis* 2007, **27**:3-12.
4. Osterreicher CH, Stickel F, Brenner DA: **Genomics of liver fibrosis and cirrhosis.** *Semin Liver Dis* 2007, **27**:28-43.
5. Juran BD, Lazaridis KN: **Genomics and complex liver disease: Challenges and opportunities.** *Hepatology* 2006, **44**:1380-90.
6. Grant A, Neuberger J: **Guidelines on the use of liver biopsy in clinical practice. British Society of Gastroenterology.** *Gut* 1999, **45**(Suppl 4): IV1-IV11.
7. McGill DB, Rakela J, Zinsmeister AR, Ott BJ: **A 21-year experience with major hemorrhage after percutaneous liver biopsy.** *Gastroenterology* 1990, **99**:1396-400.

8. Piccinino F, Sagnelli E, Pasquale G, Giusti G: **Complications following percutaneous liver biopsy. A multicentre retrospective study on 68,276 biopsies.** *J Hepatol* 1986, **2**:165-73.
9. Regev A, Berho M, Jeffers LJ, Milikowski C, Molina EG, Pyrsopoulos NT, Feng ZZ, Reddy KR, Schiff ER: **Sampling error and intraobserver variation in liver biopsy in patients with chronic HCV infection.** *Am J Gastroenterol* 2002, **97**:2614-8.
10. Siddique I, El-Naga HA, Madda JP, Memon A, Hasan F: **Sampling variability on percutaneous liver biopsy in patients with chronic hepatitis C virus infection.** *Scand J Gastroenterol* 2003, **38**:427-32.
11. Ratziu V, Charlotte F, Heurtier A, Gombert S, Giral P, Bruckert E, Grimaldi A, Capron F, Poynard T: **Sampling variability of liver biopsy in nonalcoholic fatty liver disease.** *Gastroenterology* 2005, **128**:1898-906.
12. Cox LA, Schlabritz-Loutsevitch N, Hubbard GB, Nijland MJ, McDonald TJ, Nathanielsz PW: **Gene expression profile differences in left and right liver lobes from mid-gestation fetal baboons: a cautionary tale.** *J Physiol* 2006, **572**:59-66.
13. Skripenova S, Trainer TD, Krawitt EL, Blaszyk H: **Variability of grade and stage in simultaneous paired liver biopsies in patients with hepatitis C.** *J Clin Pathol* 2007, **60**:321-4.
14. Bedossa P, Dargere D, Paradis V: **Sampling variability of liver fibrosis in chronic hepatitis C.** *Hepatology* 2003, **38**:1449-57.
15. Boess F, Kamber M, Romer S, Gasser R, Muller D, Albertini S, Suter L: **Gene expression in two hepatic cell lines, cultured primary hepatocytes, and liver slices compared to the in vivo liver gene expression in rats: possible implications for toxicogenomics use of in vitro systems.** *Toxicol Sci* 2003, **73**:386-402.
16. Wai CT, Greenson JK, Fontana RJ, Kalbfleisch JD, Marrero JA, Conjeevaram HS, Lok AS: **A simple noninvasive index can predict both significant fibrosis and cirrhosis in patients with chronic hepatitis C.** *Hepatology* 2003, **38**:518-26.
17. Martinez SM, Crespo G, Navasa M, Forns X: **Noninvasive assessment of liver fibrosis.** *Hepatology* 2011, **53**:325-35.
18. Imbert-Bismut F, Ratziu V, Pieroni L, Charlotte F, Benhamou Y, Poynard T: **Biochemical markers of liver fibrosis in patients with hepatitis C virus infection: a prospective study.** *Lancet* 2001, **357**:1069-75.
19. Poynard T, Imbert-Bismut F, Munteanu M, Messous D, Myers RP, Thabut D, Ratziu V, Mercadier A, Benhamou Y, Hainque B: **Overview of the diagnostic value of biochemical markers of liver fibrosis (FibroTest, HCV FibroSure) and necrosis (ActiTest) in patients with chronic hepatitis C.** *Comp Hepatol* 2004, **3**:8.
20. Halfon P, Imbert-Bismut F, Messous D, Antoniotti G, Benchetrit D, Cart-Lamy P, Delaporte G, Doutheau D, Klump T, Sala M, Thibaud D, Trepo E, Thabut D, Myers RP, Poynard T: **A prospective assessment of the inter-laboratory variability of biochemical markers of fibrosis (FibroTest) and activity (ActiTest) in patients with chronic liver disease.** *Comp Hepatol* 2002, **1**:3.
21. Smith JO, Sterling RK: **Systematic review: non-invasive methods of fibrosis analysis in chronic hepatitis C.** *Aliment Pharmacol Ther* 2009, **30**:557-76.
22. Rockey DC, Bissell DM: **Noninvasive measures of liver fibrosis.** *Hepatology* 2006, **43**:S113-20.
23. Shaheen AA, Wan AF, Myers RP: **FibroTest and FibroScan for the prediction of hepatitis C-related fibrosis: a systematic review of diagnostic test accuracy.** *Am J Gastroenterol* 2007, **102**:2589-600.
24. Trinchet JC, Hartmann DJ, Pateron D, Laarif M, Callard P, Ville G, Beaugrand M: **Serum type I collagen and N-terminal peptide of type III procollagen in chronic hepatitis. Relationship to liver histology and conventional liver tests.** *J Hepatol* 1991, **12**:139-44.
25. Teare JP, Sherman D, Greenfield SM, Simpson J, Bray G, Catterall AP, Murray-Lyon IM, Peters TJ, Williams R, Thompson RP: **Comparison of serum procollagen III peptide concentrations and PGA index for assessment of hepatic fibrosis.** *Lancet* 1993, **342**:895-8.
26. Gallorini A, Plebani M, Pontisso P, Chemello L, Masiero M, Mantovani G, Alberti A: **Serum markers of hepatic fibrogenesis in chronic hepatitis type C treated with alfa-2A interferon.** *Liver* 1994, **14**:257-64.
27. Yabu K, Kiyosawa K, Mori H, Matsumoto A, Yoshizawa K, Tanaka E, Furuta S: **Serum collagen type IV for the assessment of fibrosis and resistance to interferon therapy in chronic hepatitis C.** *Scand J Gastroenterol* 1994, **29**:474-9.
28. Walsh KM, Timms P, Campbell S, MacSween RN, Morris AJ: **Plasma levels of matrix metalloproteinase-2 (MMP-2) and tissue inhibitors of metalloproteinases -1 and -2 (TIMP-1 and TIMP-2) as noninvasive markers of liver disease in chronic hepatitis C: comparison using ROC analysis.** *Dig Dis Sci* 1999, **44**:624-30.
29. Boeker KH, Haberkorn CI, Michels D, Flemming P, Manns MP, Lichtinghagen R: **Diagnostic potential of circulating TIMP-1 and MMP-2 as markers of liver fibrosis in patients with chronic hepatitis C.** *Clin Chim Acta* 2002, **316**:71-81.
30. Cheung KJ, Tilleman K, Deforce D, Colle I, Van Vlierberghe H: **The HCV serum proteome: a search for fibrosis protein markers.** *J Viral Hepat* 2009, **16**:418-29.
31. Morra R, Munteanu M, Bedossa P, Dargere D, Janneau JL, Paradis V, Ratziu V, Charlotte F, Thibault V, Imbert-Bismut F, Poynard T: **Diagnostic value of serum protein profiling by SELDI-TOF ProteinChip compared with a biochemical marker, FibroTest, for the diagnosis of advanced fibrosis in patients with chronic hepatitis C.** *Aliment Pharmacol Ther* 2007, **26**:847-58.
32. Gressner OA, Weiskirchen R, Gressner AM: **Biomarkers of liver fibrosis: clinical translation of molecular pathogenesis or based on liver-dependent malfunction tests.** *Clin Chim Acta* 2007, **381**:107-13.
33. Gobel T, Vorderwulbecke S, Hauck K, Fey H, Haussinger D, Erhardt A: **New multi protein patterns differentiate liver fibrosis stages and hepatocellular carcinoma in chronic hepatitis C serum samples.** *World J Gastroenterol* 2006, **12**:7604-12.
34. Mas VR, Fisher RA, Archer KJ, Maluf DG: **Proteomics and liver fibrosis: identifying markers of fibrogenesis.** *Expert Rev Proteomics* 2009, **6**:421-31.
35. Kuramitsu Y, Nakamura K: **Current progress in proteomic study of hepatitis C virus-related human hepatocellular carcinoma.** *Expert Rev Proteomics* 2005, **2**:589-601.
36. Chen G, Gharib TG, Huang CC, Taylor JM, Misek DE, Kardia SL, Giordano TJ, Iannettoni MD, Orringer MB, Hanash SM, Beer DG: **Discordant protein and mRNA expression in lung adenocarcinomas.** *Mol Cell Proteomics* 2002, **1**:304-13.
37. Tian Q, Stepaniants SB, Mao M, Weng L, Feetham MC, Doyle MJ, Yi EC, Dai H, Thorsson V, Eng J, Goodlett D, Berger JP, Gunter B, Linseley PS, Stoughton RB, Aebersold R, Collins SJ, Hanlon WA, Hood LE: **Integrated genomic and proteomic analyses of gene expression in Mammalian cells.** *Mol Cell Proteomics* 2004, **3**:960-9.
38. Pollack JR, Perou CM, Alizadeh AA, Eisen MB, Pergamenschikov A, Williams CF, Jeffrey SS, Botstein D, Brown PO: **Genome-wide analysis of DNA copy-number changes using cDNA microarrays.** *Nat Genet* 1999, **23**:41-6.
39. Sellick GS, Longman C, Tolmie J, Newbury-Ecob R, Geenhalgh L, Hughes S, Whiteford M, Garrett C, Houlston RS: **Genomewide linkage searches for Mendelian disease loci can be efficiently conducted using high-density SNP genotyping arrays.** *Nucleic Acids Res* 2004, **32**:e164.
40. Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R, Ghandour G, Perkins N, Winchester E, Spencer J, Kruglyak L, Stein L, Hsie L, Topaloglou T, Hubbell E, Robinson E, Mittmann M, Morris MS, Shen N, Kilburn D, Rioux J, Nusbaum C, Rozen S, Hudson TJ, Lipshutz R, Chee M, Lander ES: **Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome.** *Science* 1998, **280**:1077-82.
41. Johnson JM, Castle J, Garrett-Engele P, Kan Z, Loerch PM, Armour CD, Santos R, Schadt EE, Stoughton R, Shoemaker DD: **Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays.** *Science* 2003, **302**:2141-4.
42. Yamamoto M, Wakatsuki T, Hada A, Ryo A: **Use of serial analysis of gene expression (SAGE) technology.** *J Immunol Methods* 2001, **250**:45-66.
43. Ruijter JM, Van Kampen AH, Baas F: **Statistical evaluation of SAGE libraries: consequences for experimental design.** *Physiol Genomics* 2002, **11**:37-44.
44. Aebersold R, Mann M: **Mass spectrometry-based proteomics.** *Nature* 2003, **422**:198-207.
45. Capelo JL, Carreira R, Diniz M, Fernandes L, Galesio M, Lodeiro C, Santos HM, Vale G: **Overview on modern approaches to speed up protein identification workflows relying on enzymatic cleavage and mass spectrometry-based techniques.** *Anal Chim Acta* 2009, **650**:151-9.
46. Tonge R, Shaw J, Middleton B, Rowlinson R, Rayner S, Young J, Pognan F, Hawkins E, Currie I, Davison M: **Validation and development of fluorescence two-dimensional differential gel electrophoresis proteomics technology.** *Proteomics* 2001, **1**:377-96.

47. Unlu M, Morgan ME, Minden JS: **Difference gel electrophoresis: a single gel method for detecting changes in protein extracts.** *Electrophoresis* 1997, **18**:2071-7.
48. Schwartz JC, Senko MW, Syka JE: **A two-dimensional quadrupole ion trap mass spectrometer.** *J Am Soc Mass Spectrom* 2002, **13**:659-69.
49. Marshall AG, Hendrickson CL, Jackson GS: **Fourier transform ion cyclotron resonance mass spectrometry: a primer.** *Mass Spectrom Rev* 1998, **17**:1-35.
50. Hardman M, Makarov AA: **Interfacing the orbitrap mass analyzer to an electrospray ion source.** *Anal Chem* 2003, **75**:1699-705.
51. Hu Q, Noll RJ, Li H, Makarov A, Hardman M, Graham Cooks R: **The Orbitrap: a new mass spectrometer.** *J Mass Spectrom* 2005, **40**:430-43.
52. Hutchens TA: **Automaticity and reading in learning-disabled college students.** *Ann N Y Acad Sci* 1993, **682**:357-8.
53. Issaq HJ, Veenstra TD, Conrads TP, Felschow D: **The SELDI-TOF MS approach to proteomics: protein profiling and biomarker identification.** *Biochem Biophys Res Commun* 2002, **292**:587-92.
54. De Bock M, de Seny D, Meuwis MA, Chapelle JP, Louis E, Malaise M, Merville MP, Fillet M: **Challenges for biomarker discovery in body fluids using SELDI-TOF-MS.** *J Biomed Biotechnol* 2010, **2010**:906082.
55. Domon B, Aebersold R: **Mass spectrometry and protein analysis.** *Science* 2006, **312**:212-7.
56. MacBeath G, Schreiber SL: **Printing proteins as microarrays for high-throughput function determination.** *Science* 2000, **289**:1760-3.
57. Low TY, Leow CK, Salto-Tellez M, Chung MC: **A proteomic analysis of thioacetamide-induced hepatotoxicity and cirrhosis in rat livers.** *Proteomics* 2004, **4**:3960-74.
58. Henkel C, Roderfeld M, Weiskirchen R, Berres ML, Hillebrandt S, Lammert F, Meyer HE, Stuhler K, Graf J, Roeb E: **Changes of the hepatic proteome in murine models for toxically induced fibrogenesis and sclerosing cholangitis.** *Proteomics* 2006, **6**:6538-48.
59. Shepard BD, Tuma DJ, Tuma PL: **Chronic ethanol consumption induces global hepatic protein hyperacetylation.** *Alcohol Clin Exp Res* 2010, **34**:280-91.
60. Andringa KK, King AL, Eccleston HB, Mantena SK, Landar A, Jhala NC, Dickinson DA, Squadrito GL, Bailey SM: **Analysis of the liver mitochondrial proteome in response to ethanol and S-adenosylmethionine treatments: novel molecular targets of disease and hepatoprotection.** *Am J Physiol Gastrointest Liver Physiol* 2010, **298**:G732-45.
61. Kharbanda KK, Vigneswara V, McVicker BL, Newlaczyl AU, Bailey K, Tuma D, Ray DE, Carter WG: **Proteomics reveal a concerted upregulation of methionine metabolic pathway enzymes, and downregulation of carbonic anhydrase-III, in betaine supplemented ethanol-fed rats.** *Biochem Biophys Res Commun* 2009, **381**:523-7.
62. Liu EH, Chen MF, Yeh TS, Ho YP, Wu RC, Chen TC, Jan YY, Pan TL: **A useful model to audit liver resolution from cirrhosis in rats using functional proteomics.** *J Surg Res* 2007, **138**:214-23.
63. Kirpich IA, Gobejishvili LN, Bon Homme M, Waigel S, Cave M, Arteel G, Barve SS, McClain CJ, Deaciuc IV: **Integrated hepatic transcriptome and proteome analysis of mice with high-fat diet-induced nonalcoholic fatty liver disease.** *J Nutr Biochem* 2011, **22**:38-45.
64. Zhang X, Yang J, Guo Y, Ye H, Yu C, Xu C, Xu L, Wu S, Sun W, Wei H, Gao X, Zhu Y, Qian X, Jiang Y, Li Y, He F: **Functional proteomic analysis of nonalcoholic fatty liver disease in rat models: enoyl-coenzyme a hydratase down-regulation exacerbates hepatic steatosis.** *Hepatology* 2010, **51**:1190-9.
65. Kristensen DB, Kawada N, Imamura K, Miyamoto Y, Tateno C, Seki S, Kuroki T, Yoshizato K: **Proteome analysis of rat hepatic stellate cells.** *Hepatology* 2000, **32**:268-77.
66. Kawada N, Kristensen DB, Asahina K, Nakatani K, Minamiyama Y, Seki S, Yoshizato K: **Characterization of a stellate cell activation-associated protein (STAP) with peroxidase activity found in rat hepatic stellate cells.** *J Biol Chem* 2001, **276**:25318-23.
67. Deng X, Liang J, Lin ZX, Wu FS, Zhang YP, Zhang ZW: **Natural taurine promotes apoptosis of human hepatic stellate cells in proteomics analysis.** *World J Gastroenterol* 2010, **16**:1916-23.
68. Yokoyama Y, Terai S, Ishikawa T, Aoyama K, Urata Y, Marumoto Y, Nishina H, Nakamura K, Okita K, Sakaida I: **Proteomic analysis of serum marker proteins in recipient mice with liver cirrhosis after bone marrow cell transplantation.** *Proteomics* 2006, **6**:2564-70.
69. Shimada T, Nakanishi T, Toyama A, Yamauchi S, Kanzaki A, Fujiwake H, Sato TA, Ikegawa M: **Potential implications for monitoring serum bile acid profiles in circulation with serum proteome for carbon tetrachloride-induced liver injury/regeneration model in mice.** *J Proteome Res* 2010, **9**:4490-500.
70. Smyth R, Lane CS, Ashiq R, Turton JA, Clarke CJ, Dare TO, York MJ, Griffiths W, Munday MR: **Proteomic investigation of urinary markers of carbon-tetrachloride-induced hepatic fibrosis in the Hanover Wistar rat.** *Cell Biol Toxicol* 2009, **25**:499-512.
71. Walters KA, Katze MG: **Using high-throughput genomics to study hepatitis C: what determines the outcome of infection?** *Antiviral Res* 2009, **81**:198-208.
72. Diamond DL, Jacobs JM, Paeper B, Proll SC, Gritsenko MA, Carithers RL Jr, Larson AM, Yeh MM, Camp DG, Smith RD, Katze MG: **Proteomic profiling of human liver biopsies: hepatitis C virus-induced fibrosis and mitochondrial dysfunction.** *Hepatology* 2007, **46**:649-57.
73. Molleken C, Sitek B, Henkel C, Poschmann G, Sipos B, Wiese S, Warscheid B, Broelsch C, Reiser M, Friedman SL, Tornoe I, Schlosser A, Kloppel G, Schmiegel W, Meyer HE, Holmskov U, Stuhler K: **Detection of novel biomarkers of liver cirrhosis by proteomic analysis.** *Hepatology* 2009, **49**:1257-66.
74. White IR, Patel K, Symonds WT, Dev A, Griffin P, Tsokanas N, Skehel M, Liu C, Zekry A, Cutler P, Gattu M, Rockey DC, Berrey MM, McHutchison JG: **Serum proteomic analysis focused on fibrosis in patients with hepatitis C virus infection.** *J Transl Med* 2007, **5**:33.
75. Chen L, Borozan I, Sun J, Guindi M, Fischer S, Feld J, Anand N, Heathcote J, Edwards AM, McGilvray ID: **Cell-type specific gene expression signature in liver underlies response to interferon therapy in chronic hepatitis C infection.** *Gastroenterology* 2010, **138**:1123-33, e1-3.
76. Gangadharan B, Antrobus R, Dwek RA, Zitzmann N: **Novel serum biomarker candidates for liver fibrosis in hepatitis C patients.** *Clin Chem* 2007, **53**:1792-9.
77. Suzman DL, McLaughlin M, Hu Z, Kleiner DE, Wood B, Lempicki RA, Mican JM, Suffredini A, Masur H, Polis MA, Kottilil S: **Identification of novel markers for liver fibrosis in HIV/hepatitis C virus coinfected individuals using genomics-based approach.** *AIDS* 2008, **22**:1433-9.
78. Yang L, Rudser KD, Higgins L, Rosen HR, Zaman A, Corless CL, David L, Gourley GR: **Novel biomarker candidates to predict hepatic fibrosis in hepatitis c identified by serum proteomics.** *Dig Dis Sci* 2011, **56**:3305-15.
79. Corradi F, Piscaglia F, Flori S, D'Errico-Grigioni A, Vasuri F, Tame MR, Andreone P, Boni P, Gianstefani A, Bolondi L: **Assessment of liver fibrosis in transplant recipients with recurrent HCV infection: usefulness of transient elastography.** *Dig Liver Dis* 2009, **41**:217-25.
80. Cheung KJ, Libbrecht L, Tilleman K, Deforce D, Colle I, Van Vlierberghe H: **Galectin-3-binding protein: a serological and histological assessment in accordance with hepatitis C-related liver fibrosis.** *Eur J Gastroenterol Hepatol* 2010, **22**:1066-73.
81. Huang H, Shiffman ML, Friedman S, Venkatesh R, Bzowej N, Abar OT, Rowland CM, Catanese JJ, Leong DU, Sninsky JJ, Layden TJ, Wright TL, White T, Cheung RC: **A 7 gene signature identifies the risk of developing cirrhosis in patients with chronic hepatitis C.** *Hepatology* 2007, **46**:297-306.
82. Marcolongo M, Young B, Dal Pero F, Fattovich G, Peraro L, Guido M, Sebastiani G, Palu G, Alberti A: **A seven-gene signature (cirrhosis risk score) predicts liver fibrosis progression in patients with initially mild chronic hepatitis C.** *Hepatology* 2009, **50**:1038-44.
83. Pradat P, Trepo E, Potthoff A, Bakshi R, Young B, Trepo C, Lagier R, Moreno C, Lemmers A, Gustot T, Degre D, Adler M, Wedemeyer H: **The cirrhosis risk score predicts liver fibrosis progression in patients with initially mild chronic hepatitis C.** *Hepatology* 2010, **51**:356-7.
84. Narayan R, Gangadharan B, Hantz O, Antrobus R, Garcia A, Dwek RA, Zitzmann N: **Proteomic analysis of HepaRG cells: a novel cell line that supports hepatitis B virus infection.** *J Proteome Res* 2009, **8**:118-22.
85. Dienstag JL, McHutchison JG: **American Gastroenterological Association medical position statement on the management of hepatitis C.** *Gastroenterology* 2006, **130**:225-30.
86. Asselah T, Bieche I, Narguet S, Sabbagh A, Laurendeau I, Ripault MP, Boyer N, Martinot-Peignoux M, Valla D, Vidaud M, Marcellin P: **Liver gene expression signature to predict response to pegylated interferon plus ribavirin combination therapy in patients with chronic hepatitis C.** *Gut* 2008, **57**:516-24.
87. Hayashida K, Daiba A, Sakai A, Tanaka T, Kaji K, Inaba N, Ando S, Kajiyama N, Terasaki H, Abe A, Ogasawara M, Kohara M, Harada M, Okanoue T, Ito S,

Kaneko S: **Pretreatment prediction of interferon-alfa efficacy in chronic hepatitis C patients.** *Clin Gastroenterol Hepatol* 2005, **3**:1253-9.
88. Feld JJ, Nanda S, Huang Y, Chen W, Cam M, Pusek SN, Schweigler LM, Theodore D, Zacks SL, Liang TJ, Fried MW: **Hepatic gene expression during treatment with peginterferon and ribavirin: Identifying molecular pathways for treatment response.** *Hepatology* 2007, **46**:1548-63.
89. Suppiah V, Moldovan M, Ahlenstiel G, Berg T, Weltman M, Abate ML, Bassendine M, Spengler U, Dore GJ, Powell E, Riordan S, Sheridan D, Smedile A, Fragomeli V, Muller T, Bahlo M, Stewart GJ, Booth DR, George J: **IL28B is associated with response to chronic hepatitis C interferon-alpha and ribavirin therapy.** *Nat Genet* 2009, **41**:1100-4.
90. Tanaka Y, Nishida N, Sugiyama M, Kurosaki M, Matsuura K, Sakamoto N, Nakagawa M, Korenaga M, Hino K, Hige S, Ito Y, Mita E, Tanaka E, Mochida S, Murawaki Y, Honda M, Sakai A, Hiasa Y, Nishiguchi S, Koike A, Sakaida I, Imamura M, Ito K, Yano K, Masaki N, Sugauchi F, Izumi N, Tokunaga K, Mizokami M: **Genome-wide association of IL28B with response to pegylated interferon-alpha and ribavirin therapy for chronic hepatitis C.** *Nat Genet* 2009, **41**:1105-9.
91. Thompson AJ, Fellay J, Patel K, Tillmann HL, Naggie S, Ge D, Urban TJ, Shianna KV, Muir AJ, Fried MW, Afdhal NH, Goldstein DB, McHutchison JG: **Variants in the ITPA gene protect against ribavirin-induced hemolytic anemia and decrease the need for ribavirin dose reduction.** *Gastroenterology* 2010, **139**:1181-9.
92. Sarasin-Filipowicz M, Oakeley EJ, Duong FH, Christen V, Terracciano L, Filipowicz W, Heim MH: **Interferon signaling and treatment outcome in chronic hepatitis C.** *Proc Natl Acad Sci USA* 2008, **105**:7034-9.
93. Devitt EJ, Power KA, Lawless MW, Browne JA, Gaora PO, Gallagher WM, Crowe J: **Early proteomic analysis may allow noninvasive identification of hepatitis C response to treatment with pegylated interferon alpha-2b and ribavirin.** *Eur J Gastroenterol Hepatol* 2011, **23**:177-83.
94. Sun W, Zhong F, Zhi L, Zhou G, He F: **Systematic -omics analysis of HBV-associated liver diseases.** *Cancer Lett* 2009, **286**:89-95.
95. Zhang L, Peng X, Zhang Z, Feng Y, Jia X, Shi Y, Yang H, Zhang X, Liu L, Yin L, Yuan Z: **Subcellular proteome analysis unraveled annexin A2 related to immune liver fibrosis.** *J Cell Biochem* 2010, **110**:219-28.
96. Lu Y, Liu J, Lin C, Wang H, Jiang Y, Wang J, Yang P, He F: **Peroxiredoxin 2: a potential biomarker for early diagnosis of hepatitis B virus related liver fibrosis identified by proteomic analysis of the plasma.** *BMC Gastroenterol* 2010, **10**:115.
97. Marrocco C, Rinalducci S, Mohamadkhani A, D'Amici GM, Zolla L: **Plasma gelsolin protein: a candidate biomarker for hepatitis B-associated liver cirrhosis identified by proteomic approach.** *Blood Transfus* 2010, **8(Suppl 3)**:s105-12.
98. Mohamadkhani A, Jazii FR, Sayehmiri K, Jafari-Nejad S, Montaser-Kouhsari L, Poustchi H, Montazeri G: **Plasma myeloperoxidase activity and apolipoprotein A-1 expression in chronic hepatitis B patients.** *Arch Iran Med* 2011, **14**:254-8.
99. Poon TC, Hui AY, Chan HL, Ang IL, Chow SM, Wong N, Sung JJ: **Prediction of liver fibrosis and cirrhosis in chronic hepatitis B infection by serum proteomic fingerprinting: a pilot study.** *Clin Chem* 2005, **51**:328-35.
100. Myers RP, Tainturier MH, Ratziu V, Piton A, Thibault V, Imbert-Bismut F, Messous D, Charlotte F, Di Martino V, Benhamou Y, Poynard T: **Prediction of liver histological lesions with biochemical markers in patients with chronic hepatitis B.** *J Hepatol* 2003, **39**:222-30.
101. Wang J, Jiang D, Zhang H, Lv S, Rao H, Fei R, Wei L: **Proteome responses to stable hepatitis B virus transfection and following interferon alpha treatment in human liver cell line HepG2.** *Proteomics* 2009, **9**:1672-82.
102. Seth D, Gorrell MD, Cordoba S, McCaughan GW, Haber PS: **Intrahepatic gene expression in human alcoholic hepatitis.** *J Hepatol* 2006, **45**:306-20.
103. Younossi ZM, Baranova A, Ziegler K, Del Giacco L, Schlauch K, Born TL, Elariny H, Gorreta F, VanMeter A, Younoszai A, Ong JP, Goodman Z, Chandhoke V: **A genomic and proteomic study of the spectrum of nonalcoholic fatty liver disease.** *Hepatology* 2005, **42**:665-74.
104. Bell LN, Theodorakis JL, Vuppalanchi R, Saxena R, Bemis KG, Wang M, Chalasani N: **Serum proteomics and biomarker discovery across the spectrum of nonalcoholic fatty liver disease.** *Hepatology* 2010, **51**:111-20.
105. Ulukaya E, Yilmaz Y, Moshkovskii S, Karpova M, Pyatnitskiy M, Atug O, Dolar E: **Proteomic analysis of serum in patients with non-alcoholic steatohepatitis using matrix-assisted laser desorption ionization time-of-flight mass spectrometry.** *Scand J Gastroenterol* 2009, **44**:1471-6.
106. Calvert VS, Collantes R, Elariny H, Afendy A, Baranova A, Mendoza M, Goodman Z, Liotta LA, Petricoin EF, Younossi ZM: **A systems biology approach to the pathogenesis of obesity-related nonalcoholic fatty liver disease using reverse phase protein microarrays for multiplexed cell signaling analysis.** *Hepatology* 2007, **46**:166-72.
107. Wilfred de Alwis NM, Day CP: **Genes and nonalcoholic fatty liver disease.** *Curr Diab Rep* 2008, **8**:156-63.
108. Gray J, Chattopadhyay D, Beale GS, Patman GL, Miele L, King BP, Stewart S, Hudson M, Day CP, Manas DM, Reeves HL: **A proteomic strategy to identify novel serum biomarkers for liver cirrhosis and hepatocellular cancer in individuals with fatty liver disease.** *BMC Cancer* 2009, **9**:271.
109. Charlton M: **Noninvasive indices of fibrosis in NAFLD: starting to think about a three-hit (at least) phenomenon.** *Am J Gastroenterol* 2007, **102**:409-11.
110. Juran BD, Lazaridis KN: **Genetics and genomics of primary biliary cirrhosis.** *Clin Liver Dis* 2008, **12**:349-65, ix.
111. Shackel NA, McGuinness PH, Abbott CA, Gorrell MD, McCaughan GW: **Identification of novel molecules and pathogenic pathways in primary biliary cirrhosis: cDNA array analysis of intrahepatic differential gene expression.** *Gut* 2001, **49**:565-76.
112. Li YZ, Hu CJ, Leng XM, Zhao GF, Li N, Xu Y: **Promising diagnostic biomarkers for primary biliary cirrhosis identified with magnetic beads and MALDI-TOF-MS.** *Anat Rec (Hoboken)* 2009, **292**:455-60.
113. Tahiri F, Le Naour F, Huguet S, Lai-Kuen R, Samuel D, Johanet C, Saubamea B, Tricottet V, Duclos-Vallee JC, Ballot E: **Identification of plasma membrane autoantigens in autoimmune hepatitis type 1 using a proteomics tool.** *Hepatology* 2008, **47**:937-48.
114. Song H, Sokolov M: **Analysis of protein expression and compartmentalization in retinal neurons using serial tangential sectioning of the retina.** *J Proteome Res* 2009, **8**:346-51.
115. Bowlus CL, Seeley EH, Roder J, Grigorieva J, Roder H, Caprioli RM, Gershwin M: **In situ mass spectrometry of autoimmune liver diseases.** *Cell Mol Immunol* 2011, **8**:237-42.
116. Villanueva A, Newell P, Chiang DY, Friedman SL, Llovet JM: **Genomics and signaling pathways in hepatocellular carcinoma.** *Semin Liver Dis* 2007, **27**:55-76.
117. Makridakis M, Vlahou A: **Secretome proteomics for discovery of cancer biomarkers.** *J Proteomics* 2010, **73**:2291-305.
118. Peng ZG, Fan B, Du NN, Wang YP, Gao LM, Li YH, Liu F, You XF, Han YX, Zhao ZY, Cen S, Li JR, Song DQ, Jiang JD: **Small molecular compounds that inhibit hepatitis C virus replication through destabilizing heat shock cognate 70 messenger RNA.** *Hepatology* 2010, **52**:845-53.

doi:10.1186/1755-1536-5-1
Cite this article as: Hannivoort *et al.*: Genomics and proteomics in liver fibrosis and cirrhosis. *Fibrogenesis & Tissue Repair* 2012 **5**:1.

Submit your next manuscript to BioMed Central and take full advantage of:

- **Convenient online submission**
- **Thorough peer review**
- **No space constraints or color figure charges**
- **Immediate publication on acceptance**
- **Inclusion in PubMed, CAS, Scopus and Google Scholar**
- **Research which is freely available for redistribution**

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central

Journal of the American College of Cardiology
© 2010 by the American College of Cardiology Foundation
Published by Elsevier Inc.

Vol. 56, No. 4, 2010
ISSN 0735-1097/$36.00
doi:10.1016/j.jacc.2010.02.051

STATE-OF-THE-ART PAPER

Genetics and Genomics of Stroke

Novel Approaches

Alison E. Baird, MBBS, PhD

Brooklyn, New York

Evidence for a genetic basis for stroke comes from twin and family studies and from the occurrence of a number of uncommon monogenic disorders, but the contribution of genetic factors identified for stroke so far is small. Advances in genetics and genomics may permit new insights. In recent genome-wide association studies, a number of single-nucleotide polymorphisms have been associated with specific stroke subtypes and major stroke risk factors such as diabetes and atrial fibrillation. These await replication. Studies of messenger ribonucleic acid expression have also shown promise for the development of genomic signatures for stroke classification. Stroke and coronary heart disease share some features of pathophysiology, risk, and treatment, and their genetic and genomic bases also appear to overlap. (J Am Coll Cardiol 2010;56:245–53) © 2010 by the American College of Cardiology Foundation

When the human genome was sequenced in 2003, the implications for cardiovascular and stroke medicine—in terms of the new diagnostic, therapeutic, and preventive strategies that may ultimately result—were immediately recognized (1,2). Genetic and genomic factors appear to contribute to all stages of the development of vascular disease, from predisposition, through risk and subclinical disease, to overt disease. Although the contribution of genetic factors identified for stroke is small at present, it is very possible that the sequencing of the genome may result in novel clinical applications, including the development of biomarkers for stroke.

Strokes result from focal reductions of blood flow to the brain and are second to coronary heart disease (CHD) in terms of vascular disease incidence, morbidity, and mortality. Currently, there are about 795,000 new and recurrent strokes per year in the U.S. compared with 1,350,000 new and recurrent coronary events (3). There is a substantial additional burden of asymptomatic cerebrovascular disease. About 83% of strokes are due to arterial vascular occlusion (ischemic stroke), and about 17% are due to vascular rupture (hemorrhagic stroke) (4–6).

Many scientific advances have occurred in stroke diagnosis, treatment, and prevention over the past 10 to 20 years (7,8), such as advances in neurovascular imaging and intravenous tissue plasminogen activator therapy, but important questions remain unanswered. Modifiable risk factors account for only about 60% of the population-attributable risk (PAR) for stroke (9,10), as opposed to risk factors identified for CHD, which may account for more than 90% of the attributable risk (11). The mechanisms for over 30% of ischemic strokes are not known, even after extensive workup (12–14). Stroke diagnosis is inaccurate in up to 30% of patients acutely, and it is not possible to reliably distinguish between ischemic and hemorrhagic stroke clinically. No blood-based diagnostic marker has yet been developed for stroke, unlike for acute coronary syndromes; this might be because of issues related to the blood-brain barrier. There is also no reliable way to predict which patients will develop hemorrhagic transformation in the brain after thrombolytic therapy (8).

This review begins with a description of the currently recognized genetic factors involved in stroke risk and etiology: family history, single-nucleotide polymorphisms (SNPs), and monogenic disorders. This is followed by the first results that have recently been obtained from genome-wide association and gene expression and protein expression profiling studies—new genomic methodologies that have been developed over the past decade (Online Appendix). The genetic and genomic overlap between stroke and CHD and other forms of vascular disease is becoming increasingly apparent, and some examples are provided in the final section.

Genetic Factors in Stroke Risk and Etiology

Genetic factors that are associated with increased stroke risk are a family history of stroke and a number of SNPs. A number of monogenic disorders cause stroke, either as a primary manifestation or a secondary manifestation. The contribution of these factors has been increasingly recognized over the past 10 to 20 years from findings in twin

---

From the Department of Neurology, SUNY Downstate Medical Center, Brooklyn, New York. The author has one patent relating to stroke and patents pending for intracerebral hemorrhage and vascular disease risk.

Manuscript received July 16, 2008; revised manuscript received February 2, 2010, accepted February 9, 2010.

Abbreviations and Acronyms

CADASIL = cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy

CHD = coronary heart disease

ICH = intracerebral hemorrhage

PAR = population-attributable risk

RNA = ribonucleic acid

SAH = subarachnoid hemorrhage

SNP = single-nucleotide polymorphism

studies (15) and the description of monogenic disorders such as cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy (CADASIL) (16). More recently, it was reported that Fabry disease accounts for a previously unrecognized proportion of strokes of undetermined mechanism (i.e., cryptogenic stroke) in the young (17). Evidence for a genetic basis for stroke has also come from experimental stroke models (18).

The importance of the accurate classification of stroke type and subtype needs to be appreciated, because ischemic and hemorrhagic strokes differ in their causes, pathophysiology, treatments, and outcomes (12,13) (Table 1). Careful classification of stroke type and subtype permitted the identification of genetic factors associated with CADASIL and with cerebral venous thrombosis (16,19).

Family history of stroke. In twin studies, the concordance rates for stroke are 17.7% in monozygotic twins and 3.6% in dizygotic twins (15,20). The heritability of stroke has varied in family studies, but allowing for some methodological weaknesses, it appears that a family history of stroke increases a subject's stroke risk by 2- to 3-fold (21,22). In the Framingham study, using information obtained across 3 generations, including the original and offspring cohorts, a parental history of stroke was associated with an approximately 2-fold increase in stroke risk: a paternal history of stroke was associated with a relative risk of 2.4 and a maternal history with a risk of 1.4 (21). In later studies looking at the heritability of stroke subtypes, the large artery atherosclerosis and small vessel subtypes were found to be more commonly inherited than the cardioembolic subtype (23).

Few studies have separated ischemic stroke risk from that of intracerebral hemorrhage (ICH), probably because of the smaller frequency of the ICH stroke type and the difficulty in retrospectively obtaining this information. It appears that ICH in a first-degree relative increases a subject's odds by as much as 2- to 6-fold (24). Subarachnoid hemorrhage (SAH) has a distinct clinical profile, and a family history of SAH has been associated with an increased risk for SAH of about 4-fold (25); because of this association, screening of first-degree relatives of patients with SAH is currently recommended (26). Evidence of a gene-environment interaction with smoking, another risk factor for SAH, also may exist for aneurysmal SAH (27).

Genetic polymorphisms. Polymorphisms in 2 genes increase the risk for cerebral venous thrombosis. A polymorphism in the gene coding for coagulation factor V results in a form of factor V variant, factor V Leiden, that is resistant to degradation by activated protein C, resulting in a hypercoagulable state (19). This polymorphism involves a single base pair substitution. The risk for cerebral venous thrombosis is increased almost 8-fold in subjects carrying 1 copy of the variant (28). The prothrombin G20210A polymorphism confers an almost 10-fold increased risk for cerebral venous thrombosis. The risk is further magnified with the use of the birth control pill and in the post-partum period (28). Both of these factors may be associated with a mildly increased risk for arterial ischemic stroke (19,28).

Monogenic disorders associated with stroke. These disorders are uncommon, accounting for <1% of all strokes. They are typically associated with stroke in childhood or young adulthood, with certain stroke types and subtypes, the absence of other stroke risk factors, and with specific phenotypes of the associated diseases (Table 2).

CADASIL was initially recognized in the 1990s as an adult-onset hereditary disorder of dementia and stroke (16), and its genetic and molecular basis was subsequently unraveled (Fig. 1). CADASIL is characterized by recurrent subcortical ischemic strokes starting at ages 30 to 40 years, progressive or stepwise cognitive decline, white matter changes on magnetic resonance imaging, and, in some patients, migraine with aura and mood disturbances (29,30). In addition to genetic analysis, skin biopsy is used to aid in diagnosis (31).

In other monogenic disorders—such as Fabry disease (X-linked recessive inheritance), sickle-cell disease (autosomal dominant inheritance), and mitochondrial myopathy, encephalopathy, lactic acidosis, and strokelike episodes—stroke is a secondary manifestation. In Table 2, monogenic disorders associated with stroke and vasculopathy are listed, along with disorders associated with premature atherosclerosis, arterial dissection, and monogenic forms of ICH. The reader is referred to several excellent reviews for specific details of these uncommon disorders (26,32,33).

Table 1 Major Stroke Types and Subtypes

Type	Subtype/Presumed Major Mechanism	Approximate Frequency (%)
Ischemic stroke		83
	Large artery atherosclerosis	14
	Cardioembolism	23
	Small-vessel disease	13
	Other determined	3
	Undetermined	30
Hemorrhagic stroke		17
Intracerebral hemorrhage	Hypertension	7
	Amyloid angiopathy	3
Subarachnoid hemorrhage	Intracranial aneurysm	7
Cerebral venous thrombosis		<1

The TOAST (Trial of Org 10172 in Acute Stroke Treatment) subtypes are commonly used to classify ischemic stroke (14). Common subtypes of hemorrhagic stroke are intracerebral hemorrhage and subarachnoid hemorrhage. A minority of strokes result from vascular occlusion in the cerebral venous system, termed cerebral venous thrombosis (4–6).

Table 2 Monogenic Disorders Associated With Stroke

Disorder	Causative Gene/Chromosome Location	Inheritance Pattern
Small-vessel vasculopathy		
CADASIL	Notch 3 gene	AD
CARASIL	Unknown	AR
HERNS	3p21.1-p21.3	AD
Small-artery and large-artery vasculopathy		
Fabry disease	Alpha-galactosidase A gene	XL-R
Pseudoxanthoma elasticum	ATP-binding cassette subfamily C member 6 gene	AR or AD
Large-artery vasculopathy		
Moyamoya disease	3p24.2-26, 17q25	Polygenic or AR
Premature atherosclerosis		
Homocystinuria	Cystathionine beta synthase	AR
Arterial dissection		
Ehlers-Danlos syndrome type IV	Collagen type III gene	AD
Marfan syndrome	FBN1 and TGFβR2 genes	AD
Inherited cerebral amyloid angiopathies		
Hereditary cerebral hemorrhage with amyloidosis of the Dutch type	Amyloid precursor protein gene	AD
Cystatin C–related familial cerebral amyloid angiopathy	Cystatin-C gene	AD
Familial amyloid polyneuropathy	Transthyretin gene	AD
Cerebrovascular malformations		
Autosomal dominant polycystic kidney disease	PKD1, PKD2	AD
Cerebral cavernous malformations	7q21-q22, 7p13-p15, 3q25.2-q27	AD or sporadic
Hereditary hemorrhagic telangiectasia (Osler-Weber-Rendu disease)	Endoglin and activin receptor–like kinase 1 genes	AD

Further details may be found in Meschia et al. (26), Ballabio et al. (32), and Rost et al. (33).

AD = autosomal dominant; AR = autosomal recessive; ATP = adenosine triphosphate; CADASIL = cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy; CARASIL = cerebral autosomal recessive arteriopathy with subcortical infarcts and leukoencephalopathy; FBN1 = fibrillin 1; HERNS = hereditary endotheliopathy with retinopathy, nephropathy, and stroke; PKD = polycystic kidney disease; TGFβR2 = transforming growth factor, beta receptor II; XL-R = X-linked recessive.

Genomewide Association Studies

As a result of the sequencing of the human genome, genomewide association studies may permit the detection of common variant alleles (>5%) of mild to moderate effect (34,35) associated with common disorders such as stroke. It is currently believed that common stroke may result from a number of genes of modest effect that in combination exert a significant effect on stroke risk (36). Commercial arrays containing a huge number of SNPs are now available, but the power of a typical study of several thousand patients is of the order of only 10% to 25% (34,35), so it is most important that adequate sample sizes be studied and that notable findings be validated and replicated in independent cohorts. Significant amounts of time, energy, and resources were expended in linkage studies and association studies of candidate genes over the past decade (37,38); many early findings turned out to be false positives, with studies often being limited by insufficient sample sizes.

The first genomewide association study in stroke was reported from the ISGS (Ischemic Stroke Genetics Study) in 2007 (39). No genetic locus was specifically and robustly associated with stroke, but there were only about 250 patients and controls. This group went on to study the 9p21.3 region (i.e., a region on the short arm [p] of chromosome 9) that had already been associated with CHD (40,41). An association was found with stroke, although it was not determined by genotyping that this association involved the same variants (40). Another group has since described modest associations between variants in the 9p21 region (rs2383207 and rs10757274) and ischemic stroke (42). However, in another study, 6 SNPs in the 9p21 region (including rs2383207) were found to be associated with the large artery atherosclerotic subtype of ischemic stroke, independent of CHD and vascular risk factors (43). The lead SNP, rs1537378, had a pooled odds ratio for stroke of 1.21 and a point estimate of the PAR for atherosclerotic ischemic stroke of 20.1% (43).

In a case-control study reported in 2008 (44), the 4q25 region was found to be associated with the cardioembolic subtype of ischemic stroke (with an odds ratio of about 1.5). This locus had already been found to be associated with atrial fibrillation. An association, although to a lesser degree, was found with all ischemic stroke (44); this latter finding was attributed to possible undetected cases of atrial fibrillation in patients with cryptogenic strokes. Another locus for cardioembolic stroke, on 16q22, has been recently reported (45). These studies demonstrate the value of having accurate stroke subtype classification, although having a standardized methodology in multicenter studies is a major challenge. For intracranial aneurysms, initial and seemingly quite robust associations have been found with loci on 2q, 8q, and the 9p21 region (the rs10757278 variant that is also associated with CHD) (46,47).

The first prospective genomewide association study in stroke was reported in 2009 from the CHARGE (Cohorts for Heart and Aging Research in Genomic Epidemiology)

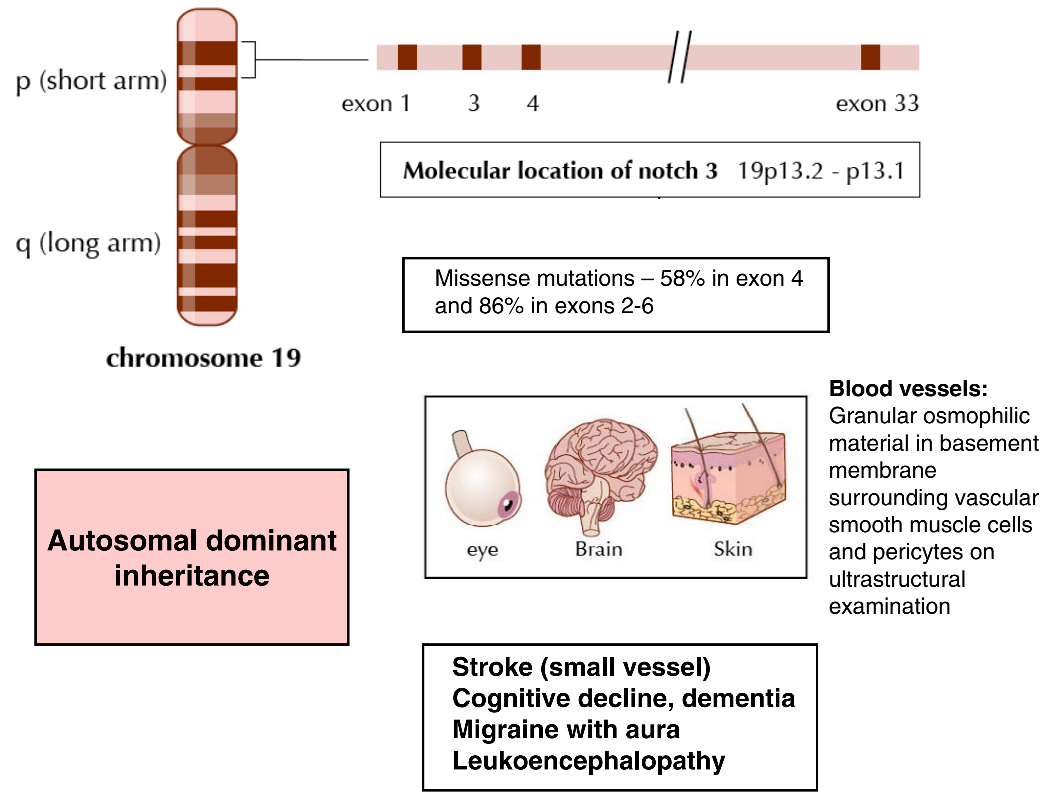

p (short arm)
q (long arm)
chromosome 19
exon 1
3
4
exon 33
Molecular location of notch 3 19p13.2 - p13.1
Missense mutations – 58% in exon 4 and 86% in exons 2-6
eye
Brain
Skin
Blood vessels:
Granular osmophilic material in basement membrane surrounding vascular smooth muscle cells and pericytes on ultrastructural examination
Autosomal dominant inheritance
Stroke (small vessel)
Cognitive decline, dementia
Migraine with aura
Leukoencephalopathy

Figure 1 **Pathophysiology of CADASIL**

Cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy (CADASIL) is an autosomal dominant disorder resulting from a number of mutations in the Notch 3 gene. This gene is located on chromosome 19, on the p (short) arm, at 19p13.2 to p13.1. This gene has 33 exons; the majority of mutations occur on exons 2 to 6 and are single-point or missense mutations. There is an accumulation of osmophilic granules in smooth muscle of blood vessels throughout the vasculature, particularly in the brain, the retina, and the skin.

Consortium (48). Consisting of 4 prospective epidemiological cohorts, this large study of nearly 19,600 subjects with 1,544 incident strokes identified 2 SNPs on chromosome 12, in the region of 12p13, although replication was obtained only for the rs12425791 SNP; the hazard ratios were 1.3 for all stroke and 1.0 for ischemic stroke. This SNP was located close to the ninjurin 2 gene, which encodes an adhesion molecule expressed in glia that shows increased expression after nerve injury. Without genotyping, it is not known if this finding is associative or causative. The International Stroke Genetics Consortium, established in 2007, is currently validating the results of the CHARGE study. The International Stroke Genetics Consortium is also collaborating with the Wellcome Trust Case Control Consortium–2 in a multistage genomewide association study of ischemic stroke. Plans are under way for the development of a similar genomewide survey in the U.S. (49).

Aside from unaccounted stroke risk, genomewide association studies may reveal genetic bases for major stroke risk factors themselves, which account for up to 60% of the PAR for stroke (9,10,50) (Fig. 2). These include hypertension (the leading modifiable stroke risk factor), which has a PAR of 30% to 40%; diabetes; smoking; atrial fibrillation; and CHD itself (with a PAR of about 4% to 5%). Loci have been associated with atrial fibrillation (4q25 and 6q22) (44,45), diabetes (loci in and around the genes cyclin-dependent kinase inhibitor 2A and cyclin-dependent kinase inhibitor 2B in the 9p21 region, among others) (51), and hypertension (serine threonine kinase 39) (52).

There has clearly been a remarkable amount of information coming to light, and a number of loci of interest are emerging, but it is too early to tell how major an impact genomewide association studies will turn out to have. In studies of other disease states, the functions of some genes have not conformed to known disease biology or not coded for proteins, as expected, but rather for ribonucleic acids (RNAs), highlighting the importance of gene expression and of micro-RNAs as future players (34–36). It is also vital that these findings be replicated and that their PARs be evaluated further; loci found for some other disease states have been of very small effect or have had little predictive value after adjustment for known risk factors (36), as may be the case for the rs10757274 polymorphism on chromosome

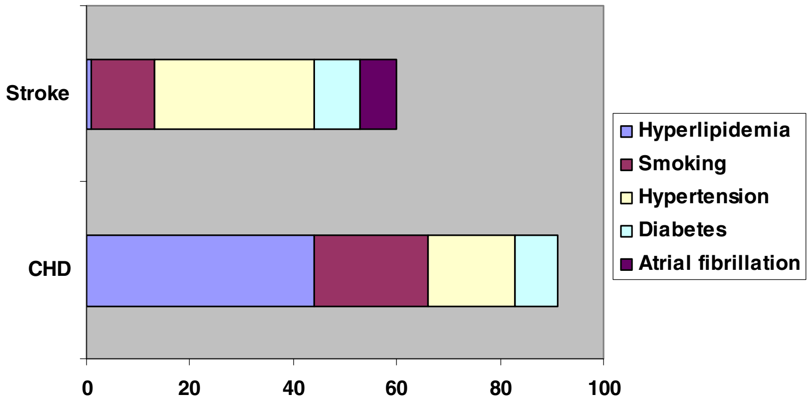


Figure 2 **Attributable Risks of Major Risk Factors for Stroke and CHD**

Vascular disorders, such as coronary heart disease (CHD) and stroke, share a number of risk factors in common. CHD also accounts for 1% to 2% of the population-attributable risk for stroke. Data were obtained from the INTERHEART (Effect of Potentially Modifiable Risk Factors Associated With Myocardial Infarction in 52 Countries) study (11) using Northern American statistics and data from current guidelines for primary prevention of stroke (50).

9p21.3 and prediction of CHD risk (53). In the future, it is likely that the focus will be on genotyping studies to look for less common variant alleles of more significant effect and in studying copy number variants as additional genetic factors associated with common disorders such as stroke.

Gene Expression Profiling, Proteomics, and Metabolomics

A number of important methodologies and technologies have been developed in tandem to those for investigating deoxyribonucleic acid variations, based on the fact that deoxyribonucleic acid is transcribed to messenger RNA, which in turn is translated into protein and metabolites. These methods can be used to investigate the entire transcriptome (gene expression profiling or RNA expression profiling), the proteome (proteomics), and the metabolome (metabolomics). They capture not only functional changes induced by genetic variations but also changes induced by environmental factors, risk factors, and the effects of interventions, and they can be used to monitor functional changes occurring over time. The interest of these methods for stroke is that they offer potential for identifying biomarkers (e.g., from the blood) that can be used in the clinic to diagnose and monitor disease and changing risk associated with mutations and treatment interventions. In genome-wide association studies, they are used to study if and to what degree genetic polymorphisms are functional.

Gene expression profiling and signatures for stroke classification. Gene expression profiling has been most widely applied in the field of oncology. Preliminary studies have been performed in carotid plaque specimens removed during endarterectomy (54).

Messenger RNA profiling of the peripheral blood has also been used to investigate a number of neurological disorders, including stroke, despite initial concerns over the specificity of signatures obtained. Blood-derived mononuclear cells (monocytes and lymphocytes) are involved in the development of atherosclerotic lesions associated with many forms of vascular disease, including stroke (55). In stroke, peripheral blood mononuclear cells are also of significance because blood-borne leukocytes, first polymorphonuclear cells and then subsequently mononuclear cells, are key players in the evolution of brain infarcts. The cells selectively migrate to and infiltrate the ischemic brain tissue. These inflammatory and immune cells may not only exacerbate ischemic reperfusion injury but may also be involved in tissue repair and remodeling after stroke (56). Changes in functional gene expression in peripheral blood mononuclear cells in brain disorders may occur in response to exposure to brain antigens (57).

Early human studies in stroke using oligonucleotide microarrays have been promising. Using peripheral blood mononuclear cells, a validated and replicated gene expression signature of ischemic stroke within the first 72 h has been identified (58). The gene expression changes seemed to reflect an adaptive response, at least in part, to the ischemic stroke, with up-regulation of genes involved in cell adhesion, inhibition of neuronal apoptosis, response to hypoxia, and vascular repair (Table 3). One example is ectonucleoside triphosphate diphosphohydrolase 1 (cluster of differentiation 39), a prominently up-regulated gene that is of interest for its cerebroprotective and cardioprotective properties. A panel of 22 genes was about 80% accurate for the detection of ischemic stroke in an external sample (58), while another group found this panel to be over 85% accurate, using whole-blood profiling (59). Signatures have since been developed for ICH (60) (prominently up-regulated genes being amphiphysin and interleukin-1 receptor 2), stroke recovery (61), and ischemic stroke subtypes (62).

Given the increasing recognition of the role of lymphocytes in stroke and vascular disease (63), and the improved charac-

Table 3 Classes of Genes Expressed in Peripheral White Blood Cells in Ischemic Stroke

Pathophysiological Class	Examples
White cell activation and differentiation (65%)	
	Toll-like receptor 2
	Ectonucleoside triphosphate diphosphohydrolase 1 (CD39)
	CD14
	Interleukin-13 receptor, alpha 1
	FcGR2A
Response to hypoxia (15%)	
	Adrenomedullin
	Leukotriene A4 hydrolase
	Cytochrome b-245
Vascular repair (10%)	
	CD36 antigen (thrombospondin receptor)
	Thrombomodulin
Response to the altered cerebral microenvironment (5%)	
	Neuronal apoptosis inhibitory protein (apoptosis inhibition)
	Catechol-O-methyl transferase (neurotransmitter degradation)
	Glutamine ligase (neurotransmitter degradation)
	Sortilin (neuronal apoptotic cell death)
	Phospholipid scramblase 1 (neurite outgrowth in neural development)
	Growth arrest specific 7 (neurite outgrowth)
Unknown (5%)	
	KIAA0146 protein

Classes of genes showing increased expression in peripheral blood mononuclear cells in the first 72 h of ischemic stroke (58). These changes were suggestive of at least a partial adaptive response to the stroke.

CD = cluster of differentiation; FcGR2A = Fc fragment of IgG, low affinity IIa, receptor.

terization of lymphocyte subpopulations, ongoing studies are investigating gene expression within leukocyte populations, along with circulating stem cells that may be involved in tissue repair and recovery after stroke. Real-time polymerase chain reaction studies are proving practical in ongoing application and validation studies. This approach offers hope for the development of a rapid diagnostic test for stroke if a suitable technology platform could be identified. This work also offers hope for the identification of gene panels for vascular disease risk and for prediction of response to treatment.

Proteomics and metabolomics. Blood-based protein biomarkers (64) could be of great value for the early detection of stroke because they could permit more accurate diagnoses to be made at the initial point of care. Currently, stroke evaluation is dependent on the patient's being taken to the hospital and undergoing a brain scan, which can delay the commencement of therapy for up to 60 to 90 min. Using a candidate protein approach, a panel of 4 proteins showed modest accuracy for the diagnosis of acute ischemic and hemorrhage stroke (65). The proteins were a marker of glial activation (S100beta), a marker of inflammation (matrix metalloproteinase-9), brain natriuretic factor, and D-dimer. Using surface-enhanced laser desorption/ionization technology, apolipoprotein C-I and apolipoprotein C-III were identified as potential plasmatic markers to distinguish between ischemic and hemorrhagic stroke (66). The results were validated using enzyme-linked immunoabsorption assays; further findings on clinical utility are awaited.

In another study that used surface-enhanced laser desorption/ionization and logical analysis-of-data classification modeling, 3 biomarkers were identified from mass spectral peaks that could detect ischemic stroke with an accuracy of 75% (67). The results were upheld in a validation sample, although the proteins were not definitively identified. The correlation and root mean absolute error of the logical analysis-of-data predictive model were superior to those of the other statistical algorithms used. Metabolomics has not yet been applied to stroke.

Pharmacogenomics. Pharmacogenomic approaches have included the evaluation and identification of genetic modifiers of clinical responses to warfarin, the statin drugs, clopidogrel, and antihypertensive therapy (68). The goals of these approaches are to personalize dosing and to minimize side effects.

Heart–Brain Linkages

Coincident with the aforementioned findings of genetic loci and gene expression changes of interest in stroke have been dramatic advances in the genetics of heart and vascular disease. A number of genetic and genomic factors already appear to overlap between stroke and heart and vascular disease, contributing to all stages of disease development: familial predisposition (21,22), vascular risk, monogenic vasculopathies, and in genomewide association studies, for example, variants in and around the 9p21 region, as previously described, and the leukotriene pathway (58,69) (Fig. 3). These

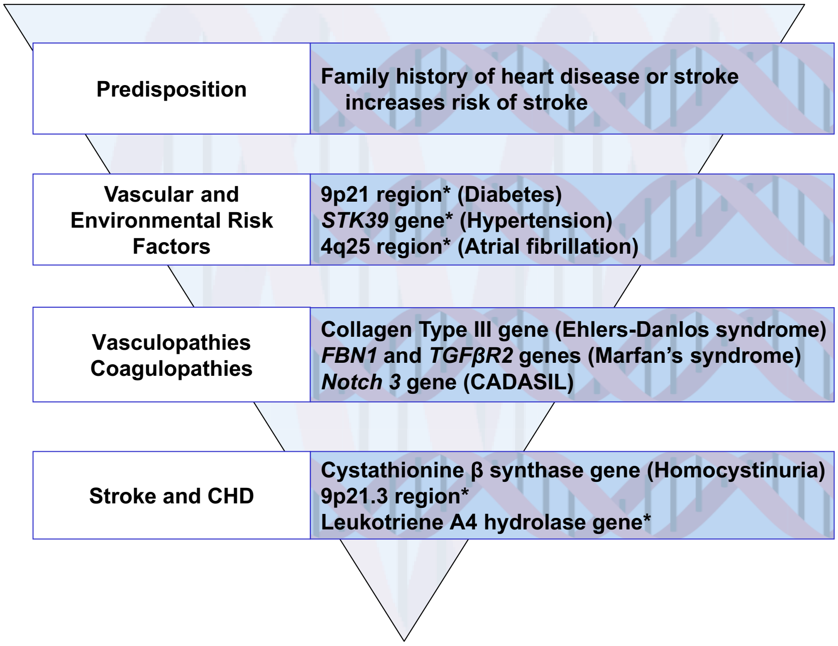

Predisposition
Family history of heart disease or stroke increases risk of stroke
Vascular and Environmental Risk Factors
9p21 region* (Diabetes)
STK39 gene* (Hypertension)
4q25 region* (Atrial fibrillation)
Vasculopathies
Coagulopathies
Collagen Type III gene (Ehlers-Danlos syndrome)
FBN1 and *TGFβR2* genes (Marfan's syndrome)
Notch 3 gene (CADASIL)
Stroke and CHD
Cystathionine β synthase gene (Homocystinuria)
9p21.3 region*
Leukotriene A4 hydrolase gene*

Figure 3 **Stroke and Heart and Vascular Disease: Genetic and Genomic Links**

Common genetic and genomic factors appear to contribute to the development of vascular disease, including familial predisposition, risk factors, subclinical disease, and overt clinical disease. Some are proven and some are putative **(*)**, having been recently reported in genomewide association studies. CADASIL = cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy; CHD = coronary heart disease; FBN1 = fibrillin 1; STK39 = serine threonine kinase 39; TGFβR2 = transforming growth factor, beta receptor II.

are deserving of further study. Furthermore, the vasculopathy of CADASIL has not been thought to involve the heart, but in 1 series, high rates of myocardial infarction were reported in patients with CADASIL (70). Higher rates of sudden unexplained death in patients with CADASIL have also been reported and have been suggested to be due to autonomic derangement, placing patients at higher risk for life-threatening arrhythmias (71).

Conclusions

At present, the contribution of genetic factors in stroke is small, involving familial predisposition, a small number of monogenic disorders such as CADASIL, and polymorphisms associated with cerebral venous thrombosis. New associations are now being reported in genomewide association studies, but these need confirmation. Future work may well focus on genotyping, looking for rarer high-risk variant alleles and studying copy number variants. It is especially important that initial results from expression profiling of peripheral blood leukocytes be further explored. Genomic methods already show particular promise for improving stroke classification. They could ultimately permit the development of biomarkers that could be valuable for the diagnosis, treatment, management, and assessment of not only stroke but many other vascular disorders.

Reprint requests and correspondence: Dr. Alison E. Baird, SUNY Downstate Medical Center, Department of Neurology, 450 Clarkson Avenue, Brooklyn, New York 11203. E-mail: alison.baird@downstate.edu.

REFERENCES

1. Ginsburg GS, Donahue MP, Newby LK. Prospects for personalized cardiovascular medicine: the impact of genomics. J Am Coll Cardiol 2005;46:1615–27.
2. Arnett DK, Baird AE, Barkley RA, et al. Relevance of genetics and genomics for prevention and treatment of cardiovascular disease: a scientific statement from the American Heart Association Council on Epidemiology and Prevention, the Stroke Council, and the Functional Genomics and Translational Biology Interdisciplinary Working Group. Circulation 2007;115:2878–901.
3. Lloyd-Jones D, Adams RJ, Brown TM, et al. Heart disease and stroke statistics—2010 update: a report from the American Heart Association. Circulation 2010;121:e46–215.
4. Incidence and Prevalence: 2006 Chart Book on Cardiovascular and Lung Diseases. Bethesda, MD: National Heart, Lung, and Blood Institute, 2006.
5. Rosamond WD, Folsom AR, Chambless LE, et al. Stroke incidence and survival among middle-aged adults: 9-year follow-up of the Atherosclerosis Risk in Communities (ARIC) cohort. Stroke 1999;30:736–43.
6. Rothwell PM, Coull AJ, Silver LE, et al. Population-based study of event-rate, incidence, case fatality, and mortality for all acute vascular events in all arterial territories (Oxford Vascular Study). Lancet 2005;366:1773–83.
7. Donnan GA, Fisher M, Macleod M, Davis SM. Stroke. Lancet 2008;371:1612–23.

8. The National Institute of Neurological Disorders and Stroke rt-PA Stroke Study Group. Tissue plasminogen activator for acute ischemic stroke. N Engl J Med 1995;333:1581–7.
9. Davis PH, Dambrosia JM, Schoenberg BS, et al. Risk factors for ischemic stroke: a prospective study in Rochester, Minnesota. Ann Neurol 1987;22:319–27.
10. Whisnant JP. Modeling of risk factors for ischemic stroke. The Willis Lecture. Stroke 1997;28:1840–4.
11. Yusuf S, Hawken S, Ounpuu S, et al. Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study. Lancet 2004;364:937–52.
12. Sacco RL, Ellenberg JH, Mohr JP, et al. Infarcts of undetermined cause: the NINCDS Stroke Data Bank. Ann Neurol 1989;25:382–90.
13. Petty GW, Brown RD Jr., Whisnant JP, Sicks JD, O'Fallon WM, Wiebers DO. Ischemic stroke subtypes: a population-based study of incidence and risk factors. Stroke 1999;30:2513–6.
14. Adams HP Jr, Bendixen BH, Kappelle LJ, et al. Classification of subtype of acute ischemic stroke. Definitions for use in a multicenter clinical trial. TOAST. Trial of Org 10172 in Acute Stroke Treatment. Stroke 1993;24:35–41.
15. Brass LM, Isaacsohn JL, Merikangas KR, Robinette CD. A study of twins and stroke. Stroke 1992;23:221–3.
16. Tournier-Lasserve E, Joutel A, Melki J, et al. Cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy maps to chromosome 19q12. Nat Genet 1993;3:256–9.
17. Rolfs A, Böttcher T, Zschiesche M, et al. Prevalence of Fabry disease in patients with cryptogenic stroke: a prospective study. Lancet 2005;366:1794–6.
18. Rubattu S, Volpe M, Kreutz R, Ganten U, Ganten D, Lindpaintner K. Chromosomal mapping of quantitative trait loci contributing to stroke in a rat model of complex human disease. Nat Genet 1996;13: 429–34.
19. Lüdemann P, Nabavi DG, Junker R, et al. Factor V Leiden mutation is a risk factor for cerebral venous thrombosis: a case-control study of 55 patients. Stroke 1998;29:2507–10.
20. Bak S, Gaist D, Sindrup SH, Skytthe A, Christensen K. Genetic liability in stroke: a long-term follow-up study of Danish twins. Stroke 2002;33:769–74.
21. Kiely DK, Wolf PA, Cupples LA, Beiser AS, Myers RH. Familial aggregation of stroke: the Framingham study. Stroke 1993;24: 1366–71.
22. Jousilahti P, Rastenyte D, Tuomilehto J, Sarti C, Vartiainen E. Parental history of cardiovascular disease and risk of stroke: a prospective follow-up of 14371 middle-aged men and women in Finland. Stroke 1997;28:1361–6.
23. Jerrard-Dunne P, Cloud G, Hassan A, Markus HS. Evaluating the genetic component of ischemic stroke subtypes: a family history study. Stroke 2003;34:1364–9.
24. Woo D, Sauerbeck LR, Kissela BM, et al. Genetic and environmental risk factors for intracerebral hemorrhage: preliminary results of a population-based study. Stroke 2002;33:1190–5.
25. Kissela BM, Sauerbeck L, Woo D, et al. Subarachnoid hemorrhage: a preventable disease with a heritable component. Stroke 2002;33: 1321–6.
26. Meschia JF, Brott TG, Brown RD Jr. Genetics of cerebrovascular disorders. Mayo Clin Proc 2005;80:122–32.
27. Woo D, Khoury J, Haverbusch MM, et al. Smoking and family history and risk of aneurysmal subarachnoid hemorrhage. Neurology 2009;72: 69–72.
28. Martinelli I, Sacchi E, Landi G, Taioli E, Duca F, Mannucci PM. High risk of cerebral-vein thrombosis in carriers of a prothrombin-gene mutation and in users of oral contraceptives. N Engl J Med 1998;338:1793–7.
29. Joutel A, Corpechot C, Ducros A, et al. Notch3 mutations in CADASIL, a hereditary adult-onset condition causing stroke and dementia. Nature 1996;383:707–10.
30. Dichgans M, Mayer M, Uttner I, et al. The phenotypic spectrum of CADASIL: clinical findings in 102 cases. Ann Neurol 1998;44:731–9.
31. Joutel A, Favrole P, Labauge P, et al. Skin biopsy immunostaining with a Notch3 monoclonal antibody for CADASIL diagnosis. Lancet 2001;358:2049–51.
32. Ballabio E, Bersano A, Bresolin N, Candelise L. Monogenic vessel diseases related to ischemic stroke: a clinical approach. J Cereb Blood Flow Metab 2007;27:1649–62.
33. Rost NS, Greenberg SM, Rosand J. The genetic architecture of intracerebral hemorrhage. Stroke 2008;39:2166–73.
34. Hardy J, Singleton A. Genomewide association studies and human disease. N Engl J Med 2009;360:1759–68.
35. Attia J, Ioannidis JP, Thakkinstian A, et al. How to use an article about genetic association: A: background concepts. JAMA 2009;301: 74–81.
36. Hunter DJ, Altshuler D, Rader DJ. From Darwin's finches to canaries in the coal mine—mining the genome for new biology. N Engl J Med 2008;358:2760–3.
37. Rosand J, Bayley N, Rost N, de Bakker PI. Many hypotheses but no replication for the association between PDE4D and stroke. Nat Genet 2006;38:1091–2.
38. Matarin M, Brown WM, Dena H, et al. Candidate gene polymorphisms for ischemic stroke. Stroke 2009;40:3436–42.
39. Matarín M, Brown WM, Scholz S, et al. A genome-wide genotyping study in patients with ischaemic stroke: initial analysis and data release. Lancet Neurol 2007;6:414–20.
40. Matarin M, Brown WM, Singleton A, Hardy JA, Meschia JF; ISGS Investigators. Whole genome analyses suggest ischemic stroke and heart disease share an association with polymorphisms on chromosome 9p21. Stroke 2008;39:1586–9.
41. McPherson R, Pertsemlidis A, Kavaslar N, et al. A common allele on chromosome 9 associated with coronary heart disease. Science 2007; 316:1488–91.
42. Smith JG, Melander O, Lovkvist H, et al. Common genetic variants on chromosome 9p21 confers risk of ischemic stroke: a large-scale genetic association study. Circ Cardiovasc Genet 2009;2:159–64.
43. Gschwendtner A, Bevan S, Cole JW, et al. Sequence variants on chromosome 9p21.3 confer risk for atherosclerotic stroke. Ann Neurol 2009;65:531–9.
44. Gretarsdottir S, Thorleifsson G, Manolescu A, et al. Risk variants for atrial fibrillation on chromosome 4q25 associate with ischemic stroke. Ann Neurol 2008;64:402–9.
45. Gudbjartsson DF, Holm H, Gretarsdottir S, et al. A sequence variant in ZFHX3 on 16q22 associates with atrial fibrillation and ischemic stroke. Nat Genet 2009;41:876–8.
46. Helgadottir A, Thorleifsson G, Magnusson KP, et al. The same sequence variant on 9p21 associates with myocardial infarction, abdominal aortic aneurysm and intracranial aneurysm. Nat Genet 2008; 40:217–24.
47. Bilguvar K, Yasuno K, Niemelä M, et al. Susceptibility loci for intracranial aneurysm in European and Japanese populations. Nat Genet 2008;40:1472–7.
48. Ikram MA, Seshadri S, Bis JC, et al. Genomewide association studies of stroke. N Engl J Med 2009;360:1718–28.
49. Meschia JF. Whole genome approaches to ischemic stroke. Stroke 2009;40 Suppl:S61–3.
50. Goldstein LB, Adams R, Alberts MJ, et al. Primary prevention of ischemic stroke: a guideline from the American Heart Association/American Stroke Association Stroke Council: cosponsored by the Atherosclerotic Peripheral Vascular Disease Interdisciplinary Working Group; Cardiovascular Nursing Council; Clinical Cardiology Council; Nutrition, Physical Activity, and Metabolism Council; and the Quality of Care and Outcomes Research Interdisciplinary Working Group. Circulation 2006;113:e873–923.
51. Diabetes Genetics Initiative of Broad Institute of Harvard and MIT, Lund University, and Novartis Institutes of BioMedical Research, Saxena R, et al. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. Science 2007;316:1331-6.
52. Wang Y, O'Connell JR, McArdle PF, et al. Whole-genome association study identifies STK39 as a hypertension susceptibility gene. Proc Natl Acad Sci U S A 2009;106:226–31.
53. Paynter NP, Chasman DI, Buring JE, Shiffman D, Cook NR, Ridker PM. Cardiovascular disease risk prediction with and without knowledge of genetic variation at chromosome 9p21.3. Ann Intern Med 2009;150:65–72.
54. Fortunato G, Di Taranto MD, Bracale UM, et al. Decreased paraoxonase-2 expression in human carotids during the progression of atherosclerosis. Arterioscler Thromb Vasc Biol 2008;28:594–600.
55. Hallenbeck JM, Hansson GK, Becker KJ. Immunology of ischemic vascular disease: plaque to attack. Trends Immunol 2005;26:550–6.
56. Becker K, Kindrick D, McCarron R, Hallenbeck J, Winn R. Adoptive transfer of myelin basic protein-tolerized splenocytes to naive animals

reduces infarct size: a role for lymphocytes in ischemic brain injury? Stroke 2003;34:1809–15.

57. Alvord EC Jr, Hsu PC, Thron R. Leukocyte sensitivity to brain fractions in neurological diseases. Arch Neurol 1974;30:296–9.
58. Moore DF, Li H, Jeffries N, et al. Using peripheral blood mononuclear cells to determine a gene expression profile of acute ischemic stroke: a pilot investigation. Circulation 2005;111:212–21.
59. Tang Y, Xu H, Du X, et al. Gene expression in blood changes rapidly in neutrophils and monocytes after ischemic stroke in humans: a microarray study. J Cereb Blood Flow Metab 2006;26:1089–102.
60. Moore DF, Goldin R, Yu H, et al. Using peripheral blood mononuclear cells to determine a gene expression profile of acute intracerebral hemorrhage. Ann Neurol 2007;62 Suppl:S8.
61. Moore DF, Wright V, Yu H, et al. Gene expression changes during recovery from ischemic stroke: a longitudinal study and data analysis. Ann Neurol 2005;58:S42–3.
62. Xu H, Tang Y, Liu DZ, et al. Gene expression in peripheral blood differs after cardioembolic compared with large-vessel atherosclerotic stroke: biomarkers for the etiology of ischemic stroke. J Cereb Blood Flow Metab 2008;28:1320–8.
63. Baird AE. The forgotten lymphocyte: immunity and stroke. Circulation 2006;113:2035–6.
64. Baird AE. Blood biologic markers of stroke: improved management, reduced cost? Curr Atheroscler Rep 2006;8:267–75.
65. Laskowitz DT, Kasner SE, Saver J, Remmel KS, Jauch EC, BRAIN Study Group. Clinical usefulness of a biomarker-based diagnostic test for acute stroke: the Biomarker Rapid Assessment in Ischemic Injury (BRAIN) study. Stroke 2009;40:77–85.
66. Allard L, Lescuyer P, Burgess J, et al. ApoC-I and ApoC-III as potential plasmatic markers to distinguish between ischemic and hemorrhagic stroke. Proteomics 2004;4:2242–51.
67. Reddy A, Wang H, Yu H, et al. Logical analysis of data (LAD) model for the early diagnosis of acute ischemic stroke. BMC Med Inform Decis Mak 2008;8:30.
68. Meschia JF. Pharmacogenetics and stroke. Stroke 2009;40:3641–5.
69. Helgadottir A, Manolescu A, Helgason A, et al. A variant of the gene encoding leukotriene A4 hydrolase confers ethnicity-specific risk of myocardial infarction. Nat Genet 2006;38:68–74.
70. Lesnik Oberstein SA, Jukema JW, Van Duinen SG, et al. Myocardial infarction in cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy (CADASIL). Medicine (Baltimore) 2003;82:251–6.
71. Rufa A, Guideri F, Acampa M, et al. Cardiac autonomic nervous system and risk of arrhythmias in cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy (CADASIL). Stroke 2007;38:276–80.

Key Words: genomics ■ stroke ■ cerebrovascular disorders ■ genetics.

APPENDIX

For a glossary of terms, please see the online version of this article.

Cancers **2013**, *5*, 1271-1305; doi:10.3390/cancers5041271

OPEN ACCESS

cancers

ISSN 2072-6694

www.mdpi.com/journal/cancers

Review

Strategies in Gene Therapy for Glioblastoma

Aneta Kwiatkowska [†], **Mohan S. Nandhu** [†], **Prajna Behera, E. Antonio Chiocca and Mariano S. Viapiano** *

Department of Neurosurgery, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA; E-Mails: akwiatkowska@partners.org (A.K.); nmohansobhana@partners.org (M.S.N.); pbehera@partners.org (P.B.); eachiocca@partners.org (E.A.C.)

[†] These authors contributed equally to this work.

* Author to whom correspondence should be addressed; E-Mail: mviapiano@partners.org; Tel.: +1-617-525-5650; Fax: +1-617-730-3050.

Received: 3 October 2013 / Accepted: 15 October 2013 / Published: 23 October 2013

Abstract: Glioblastoma (GBM) is the most aggressive form of brain cancer, with a dismal prognosis and extremely low percentage of survivors. Novel therapies are in dire need to improve the clinical management of these tumors and extend patient survival. Genetic therapies for GBM have been postulated and attempted for the past twenty years, with variable degrees of success in pre-clinical models and clinical trials. Here we review the most common approaches to treat GBM by gene therapy, including strategies to deliver tumor-suppressor genes, suicide genes, immunomodulatory cytokines to improve immune response, and conditionally-replicating oncolytic viruses. The review focuses on the strategies used for gene delivery, including the most common and widely used vehicles (*i.e.*, replicating and non-replicating viruses) as well as novel therapeutic approaches such as stem cell-mediated therapy and nanotechnologies used for gene delivery. We present an overview of these strategies, their targets, different advantages, and challenges for success. Finally, we discuss the potential of gene therapy-based strategies to effectively attack such a complex genetic target as GBM, alone or in combination with conventional therapy.

Keywords: gene therapy; oncolytic virus; stem cells; nanotechnology; gene transfer; glioma invasion; suicide gene; immunomodulation

---

1. Introduction

Glioblastoma (GBM) is the most aggressive tumor of the Central Nervous System (CNS) and its prognosis is one of the worst among all cancer types. Although the number of GBM cases is small compared to other solid tumors, population statistics still reveal a dim picture after years of research and improvement in the clinical management of this disease [1]. Of the approximately 13,000 new patients diagnosed with GBM in the US every year, almost 50% of the patients die within one year and 90% within three years following diagnosis [2], causing more years of life lost than most types of cancer [3].

GBM is therefore considered a fatal malignancy, incurable by conventional therapeutic strategies [4]. A major factor that contributes to the dismal prognosis of GBM is the physical and physiological isolation of these tumors within the CNS, which makes difficult the delivery of chemotherapeutics. In addition, the CNS is largely regarded as an immune sanctuary protected from systemic immune responses, therefore facilitating immune evasion of tumor cells and limiting the efficacy of systemic immune-boosting approaches. A third and critical factor that makes these tumors extremely difficult to eradicate is the highly invasive nature of GBM cells, which disperse along blood vessels and white matter, resulting in a disseminated disease that is impossible to completely resect [4]. Finally, the presence of a disseminated tumor stem-like cell population that supports tumor self-renewal and is particularly resistant to chemo- and radio-therapy, is another major factor underlying tumor recurrence and poor long-term survival [5].

Given the resistance of these tumors to conventional therapeutic approaches there is an urgent need to develop alternative strategies to complement or improve current approaches and improve long-term patient survival. Strategies under development include novel adjuvant chemotherapeutics to be combined with standard care, as well as novel molecularly-targeted approaches against the tumor and its microenvironment. In this review we will focus on a host of molecularly-targeted approaches collectively aggregated under the concept of *gene therapy*.

Gene "therapy" as the possibility of selecting the genetic information of organisms was first mentioned even before the identification of DNA as genetic material [6], but the formal concept of gene therapy as horizontal transfer of genetic material with the potential to treat diseases only solidified in the early 1970s [7], when technological advances allowed researchers to deliver tailored genetic material to mammalian cells. "Strict" gene therapy is based on the ability to replace a defective gene function through delivery and integration of the functional version of the gene. Therefore, gene therapy strategies have largely been developed for genetic diseases with clear dependency on a single gene deficiency [8], such as recessive enzymatic deficiencies and blood disorders.

Despite being a genetic disease, the possibility of applying strict gene therapy for cancer is less straightforward since tumors develop through multiple known and unknown genetic abnormalities. Moreover, the accumulation of mutations and evolution of the tumor's genetic makeup during malignant progression make cancers a genetic moving target that would defeat the single gene-replacement approach. Therefore, the concept of gene therapy for cancer has been widened to encompass the general delivery of therapeutic genetic material to the tumor, to kill cancer cells or enhance the immune response against them.

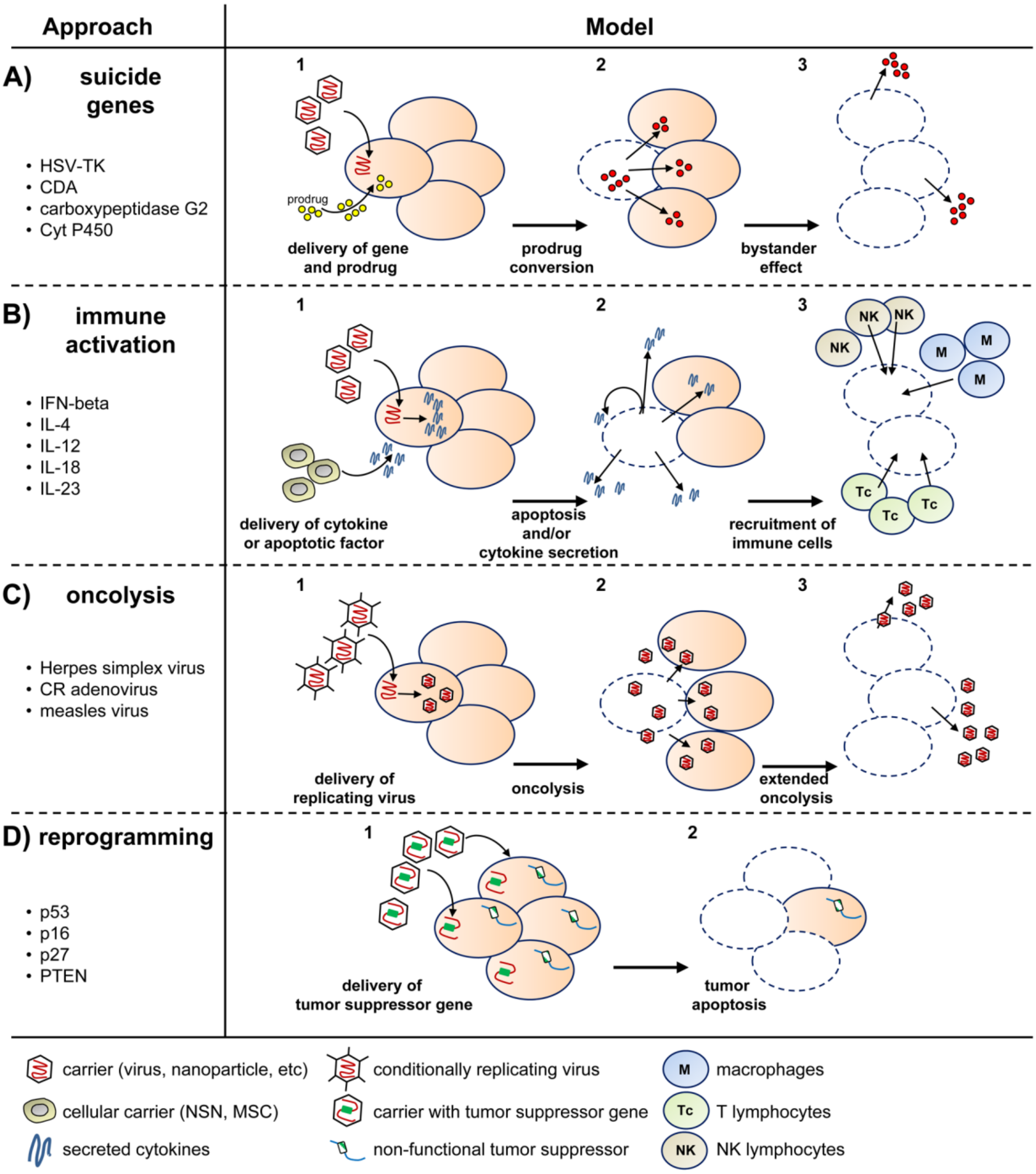
Strategies for gene therapy of cancer in general, and gliomas in particular, have been in development for the past twenty years, with a strong record of success in pre-clinical models and an increasing number of models reaching clinical trials [9] (see Table 1 for a summary of active trials). Major approaches employed for gene therapy of GBM have included: (1) delivery of suicide genes to convert prodrugs in the tumor and achieve tumor cell death; (2) delivery of cytokine genes to activate and attract immune cells against the tumor; (3) delivery of tumor-suppressor genes to reprogram tumor cells into apoptosis; and (4) delivery of conditionally-replicating viruses to specifically lyse tumor cells while sparing normal tissue. Carriers of genetic material have usually been viruses, but alternative vehicles such as stem cells, nanoparticles and liposomes, have also been extensively developed and reached the clinical stage. The following sections will describe these approaches in detail, comparing the advantages and specific challenges faced by each one. A summary of these strategies and examples of representative genes employed are shown in Figure 1.

Table 1. Active clinical trials for gene therapy of GBM. Clinical trials listed in this table are registered with *active* status (open, recruiting or ongoing) as of May 2013. Source: US National Institutes of Health [10] and Journal of Gene Medicine [11].

Country/Identifier	Model	Strategy/goals	Carrier	Phase
US/NCT00589875	AdV-TK	Suicide gene	non-replicating virus	IIa
China/CT00870181	AdV-TK	Suicide gene	non-replicating virus	II
US/NCT00634231	AdV-TK (plus radiotherapy)	Suicide gene	non-replicating virus	I
US/NCT00751270	AdV-TK (plus radiotherapy)	Suicide gene	non-replicating virus	Ib *
US/NCT00589875	AdV-TK (plus radiotherapy)	Suicide gene	non-replicating virus	IIa *
US/NCT01811992	(1) AdV-hCMV-TK and (2) AdV-hCMV-Flt3L	(1) Suicide gene (2) Immune stimulation	non-replicating virus	I
US/NCT01156584	retroviral vector (Toca-511) carrying CDA	Suicide gene and viral oncolysis	replicating virus	I/II
US/NCT01174537	New Castle Disease Virus	Viral oncolysis	replicating virus	I/II
US/NCT01301430	H-1 parvovirus (ParvOryx-01)	Viral oncolysis	replicating virus	I/II
US/NCT01491893	engineered chimeric poliovirus (PVS-RIPO)	Viral oncolysis and immune stimulation	replicating virus	I
US/NCT00390299	Engineered measles virus (MV-CEA)	Viral oncolysis and immune activation	replicating virus	I
US/NCT01582516	AdV-Delta-24-RGD delivered via CED	Viral oncolysis	replicating virus	I/II
US/NCT00805376	AdV-Delta-24-RGD-4C	Viral oncolysis	replicating virus	I
UK/UK-0050	HSV 1716	Viral oncolysis	replicating virus	II
US/NCT01172964	stem cells carrying CDA	Suicide gene	neural stem cells	Pilot

Abbreviations: AdV, adenovirus; CDA, cytosine deaminase; CED, convection-enhanced delivery; CMV, cytomegalovirus promoter; Flt3L, FMS-like tyrosine kinase 3 ligand; TK, thymidine kinase; *: E.A.C. is currently involved in these two active (non-recruiting) clinical trials.

Figure 1. Different strategies for gene therapy of GBM. (**A**) Suicide genes: GBM cells receive the suicide gene by local injection of a carrier, together with systemic delivery of a prodrug (1). The suicide gene converts the prodrug into a cytotoxic product (2) that kills the recipient cell and non-transduced bystander tumor cells (3); (**B**) Immune activation: The gene for an immunomodulatory cytokine is delivered to the tumor cells using several possible vehicles with tumor tropism (1), including viruses or stem cells. Cytokine expression increases tumor cell apoptosis (2) and activates immune cells such as macrophages, natural killer cells or T-cell lymphocytes (3); (**C**) Oncolysis: The tumor is infected with conditionally-replicating oncolytic viruses (1) that lyse the tumor cells (2–3) while sparing normal ones; (**D**) Reprogramming: Tumors receive the functional copy of a tumor suppressor gene (1), which subsequently induces cell cycle arrest or apoptosis (2).


Approach
Model
A) suicide genes
• HSV-TK
• CDA
• carboxypeptidase G2
• Cyt P450
1
2
3
prodrug
delivery of gene and prodrug
prodrug conversion
bystander effect
B) immune activation
• IFN-beta
• IL-4
• IL-12
• IL-18
• IL-23
delivery of cytokine or apoptotic factor
apoptosis and/or cytokine secretion
recruitment of immune cells
NK
M
Tc
C) oncolysis
• Herpes simplex virus
• CR adenovirus
• measles virus
delivery of replicating virus
oncolysis
extended oncolysis
D) reprogramming
• p53
• p16
• p27
• PTEN
delivery of tumor suppressor gene
tumor apoptosis
carrier (virus, nanoparticle, etc)
cellular carrier (NSN, MSC)
secreted cytokines
conditionally replicating virus
carrier with tumor suppressor gene
non-functional tumor suppressor
macrophages
T lymphocytes
NK lymphocytes

2. Virus-Based Gene Therapy of GBM

Viruses targeting mammalian cells have evolved as effective vehicles for horizontal gene transfer and have therefore been the preferred approach for gene therapy since its beginnings [7]. Moreover, the specific neurotropism of certain herpesviruses, adenoviruses and paramyxoviruses [12–16] has made them particularly valuable to target cells of the neural lineage, including malignant brain tumor cells. The first attempt to treat gliomas with a non-engineered virus was an unsuccessful study using attenuated mumps virus, conducted in 1982 [17]. A decade later, two improved strategies using engineered viruses were described almost simultaneously: the use of transduced cells delivering engineered retroviruses into the tumor stroma [18,19], and the first use of an engineered Herpes simplex virus (HSV) for selective replication in glioma cells [20]. This research marked the beginning of two of the major strategies used in glioma virotherapy: (a) targeting the tumor with replication-deficient viruses carrying conditionally-expressing suicide genes; and (b) using tumor-specific, replication-competent oncolytic viruses. The following sections describe the major strategies that have been employed for GBM gene therapy using viruses and their current pre-clinical and clinical status.

2.1. Viral Delivery of Suicide Genes

Systemic chemotherapy of tumor cells is usually limited by toxic side effects caused on dividing normal cells. Suicide gene therapy was envisioned as a way to overcome this limitation, and is based on the systemic delivery of an inactive prodrug together with tumor-specific expression of a drug-activating enzyme (the *suicide gene*) [21,22]. Suicide genes are usually absent or expressed at very low levels in mammalian cells [23] and are therefore delivered using a viral transduction system [9].

The best studied suicide gene is the HSV-derived enzyme Thymidine Kinase (HSV-TK) [9,24,25]. This enzyme catalyzes the phosphorylation of cytotoxic nucleoside analogues that can be incorporated into the DNA of actively proliferating cells, disrupting DNA replication and halting cell division. Since the prodrug nucleosides are poor substrates for mammalian TK, the toxic effect can be restricted to actively dividing cells that have been transduced with HSV-TK using non-replicating herpesvirus or adenovirus [9,23]. This strategy was first used by Ezzedine and colleagues in 1991 to demonstrate the selectivity and efficacy of HSV-TK to kill subcutaneously-implanted glioma cells upon administration of the prodrug ganciclovir [26]. Improvements of this strategy have included the use of different drug formulations to enable sustained intratumoral drug delivery [27] and the use of mutant HSV-TK versions that confer increased sensitivity to the antiviral prodrugs [25,28].

An added advantage of suicide gene therapy is the spread of cytotoxicity from the originally infected cells to neighboring neoplastic, non-infected cells, an effect known as *bystander cytotoxic effect* [22,29,30]. In the case of TK, however, the bystander effect is somewhat limited because phosphorylated nucleoside analogues do not cross the cell membrane. Instead, they must be transferred to the neighboring cells via gap-junctions or by release of apoptotic vesicles from the infected, dying cell [31–33].

In spite of the initial promise of suicide gene therapy for glioma, further evidence suggested that even HSV-TK expressing cells could become resistant to the prodrugs, therefore requiring combination of this molecularly-directed gene therapy with conventional chemo-radiotherapy [25,28,34–36].

There has also been concern about possible toxic effects, poor rate of delivery of HSV-TK to the tumor cells [37], and immune response against the delivery vehicle. For example, non-human primates treated with adenovirus-delivered HSV-TK showed dose-dependent toxicity and developed antibodies against the viral particles [38]. Chronic inflammatory symptoms (including macrophage activation and lymphocyte infiltration) were also observed in the brain of long-term surviving rats that had been implanted with intracranial gliomas and treated with adenovirus-delivered HSV-TK [39]. Despite these caveats, viral-delivered HSV-TK has proven to be a safe strategy in multiple phase I and II clinical trials [40–42] and continues to be the most common suicide gene approach in active trials (see Table 1 for active clinical trials). Side effects have been fairly minimal and the major limiting factor has not been toxicity but lack of significant improvement in efficacy against placebo. This was demonstrated in a large, multicenter, phase III clinical trial for HSV-TK (GLI328 International Study Group) that employed retrovirus-producing cells to deliver HSV-TK gene therapy in patients with newly-diagnosed GBM [37]. The trial reported a good safety profile for this adjuvant treatment, although there were no significant improvements in progression-free or overall survival. This lack of effect was largely attributed to poor distribution of the carrier and limited delivery of HSV-TK into the tumor.

Another widely studied suicide gene is the bacterial enzyme Cytosine Deaminase (CDA), which converts the prodrug 5-fluorocytosine to the toxic compound 5-fluorouracil (5-FU). 5-FU can be further converted to 5-fluorouracil triphosphate, which interferes with RNA processing, or 5-fluorouridine-5'-monophosphate, which irreversibly inhibits DNA synthesis [23]. Importantly, 5-FU can diffuse to neighboring cells and achieves bystander cytotoxic effect that does not require the presence of physical cell-cell contacts [21,43]. Further enhancement of cytotoxicity has been achieved by using an engineered bacterial CDA (bCDA-Asp^{314}Ala) with increased affinity for 5-fluorocytosine. Combination of this recombinant CDA with radiotherapy has shown significant tumor cell killing and delayed tumor growth in xenograft models of glioma [23]. Adenovirus-delivered CDA has also been combined with a second enzyme, Uracil Phosphoribosyl Transferase (UPRT), which catalyzes the conversion of 5-FU into 5-fluorourydine-5'-monophosphate. Simultaneous expression of CDA and UPRT genes has shown cooperative antitumor effects [41]. Interestingly, the sensitivity of glioma cells to the combination of CDA and UPRT plus systemic 5-fluorocytosine seemed to be p53-dependent [44], suggesting that p53 status could be used as stratification criteria for this treatment. Viral-delivered, CDA-based therapy has reached the clinical stage and a non-lytic, replicating retroviral vector (Toca-511) [45] is currently being tested in a phase I/II clinical trial to deliver the enzyme in combination with 5-fluorocytosine in patients with recurrent high-grade glioma (NIH trial NCT01156584, Table 1).

A third example of nucleoside-modifying suicide gene therapy is the use of *E. coli*-derived Purine Nucleoside Phosphorylase (PNP), which can convert non-toxic adenosine ribonucleosides (e.g., fludarabine) into toxic adenine analogs (2-fluoroadenine) that disrupt RNA processing. These metabolites can diffuse to neighboring cells, resulting in robust bystander effect in proliferating and non-proliferating cells [46,47]. Retrovirus-carried PNP has been shown to integrate in the host cell DNA, leading to long-term effect of this treatment *in vivo* [47]. An important development of this approach was the combination of herpesvirus-delivered PNP with antibiotics to remove intestinal flora that could convert the prodrug outside the tumor. This approach allowed the use of lower doses of the cytotoxic agent, enhancing chemoprotection and efficacy in a mouse model of glioma [48].

Nucleic acid-targeting gene therapies have also used transgenes coding for enzymes that generate DNA-alkylating compounds. This strategy has a considerable advantage because cytotoxicity does not depend on DNA replication or RNA expression, therefore killing both proliferating as well as quiescent glioma cells. An example of pro-alkylating suicide gene that has been tested in several solid tumor models is the bacterial enzyme carboxypeptidase G2 (CPG2). When combined with nitrogen mustards prodrugs (such as 4-[(2-chloroethyl)(2-mesyloxyethyl)amino]benzoyl-L-glutamic acid (CMDA) or ZD2767P), CPG2 yields mustard-like compounds that crosslink DNA [49]. This conversion is specific to CPG2, which has no mammalian homologues, therefore preventing non-specific generation of highly toxic alkylating compounds in off-target tissues. Engineering of this enzyme for cell-surface expression has been used to enhance the bystander effect of prodrugs that cross the cell membrane poorly and do not reach therapeutic effect with intracellular CPG2 [50]. Although CPG2 has not been tested in gliomas *in vivo*, it has been delivered to cultured glioma cells using replication-deficient adenoviruses and shown cytotoxicity comparable or higher than HSV-TK in the same cells [51].

One extensively studied pro-alkylating suicide gene therapy has used cytochrome P450 (*CYP2B1* gene), the only example of a mammalian-derived suicide gene. This cytochrome can hydroxylate the immunomodulatory prodrug cyclophosphamide (CPA), generating an alkylating phosphoramide mustard. The first demonstration of this strategy for gliomas *in vivo* used fibroblasts transduced with a replicating retrovirus carrying CYP2B1, which were injected intratumorally in mice carrying intracranial GBM xenografts [52]. Following intrathecal administration of CPA, this study demonstrated partial regression of the intracranial tumor mass and limited or absent tumor dispersion to the meningeal space. Further development of this strategy placed the CYP2B1 gene in a replicating HSV (see description of oncolytic HSV in Section 2.5), showing strong antitumor effect when CPA chemotherapy was combined with HSV-mediated tumor oncolysis [53]. A significant advantage of CPA is the inhibitory effect of this drug on the innate immune activity against HSV, therefore reducing viral clearance and enhancing both gene delivery and viral oncolysis [54].

Finally, it should be noted that the concept of suicide genes can be expanded to include additional examples of cytotoxic genes as long as they are specifically delivered to and active in the target tumor cells. One example is the potent *Pseudomonas* exotoxin A, a cytotoxin produced by *Pseudomonas aeruginosa* that disrupts protein synthesis. A truncated form of this toxin conjugated to mutated interleukin IL-13 (mhIL-13-PE, marketed under the name Cintredekin Besudotox) specifically binds and kills glioma cells while sparing normal neural cells that lack the glioma-specific receptor for the interleukin (IL13Rα2) [55]. In an elegant development of this strategy, an adenoviral vector was developed for conditional expression of mhIL-13-PE in transduced glioma cells [56]. The vector also expressed IL-4 to saturate the normal receptor of IL-13 (IL4R/IL13R), therefore preventing binding of the chimeric toxin to normal neural cells and achieving specific cytotoxicity. Inoculation of this virus in multiple models of intracranial glioma using athymic and immunocompetent mice resulted in significant tumor toxicity and increased animal survival with high proportion of long-term survivors [56].

2.2. Viral Delivery of Tumor-Suppressor Genes

Tumor suppressor genes are major regulators of DNA repair, cell proliferation, and apoptosis. Deletions and inactivating mutations in those genes are common in all cancers, including gliomas. In particular, mutations in three pathways containing tumor suppressors are commonly associated with high-grade human gliomas: p53/MDM2, p16/Rb, and PTEN [57]. Based on this feature of brain tumor biology, viral strategies have been designed in an attempt to reprogram tumor cells by restoring tumor suppressor activity in cells carrying inactivating mutations in those genes.

p53 is considered one of the most critical mediators of growth arrest and apoptosis in response to DNA damage, hypoxia, and growth factor withdrawal. This is one of the most frequently mutated genes in gliomas, being inactivated in about 30% of primary and 65% of secondary GBMs [58]. Tumor suppressor therapy using p53 in glioma cells was first attempted by delivering this gene, under control of the potent CMV promoter, using non-replicating adenovirus serotype 5 (Ad5CMV-p53). Restoration of the functional gene induced robust apoptosis of the infected cells *in vitro* and reduced tumorigenesis *in vivo* [59]. However, inoculation of the same vector in GBM-bearing mice was insufficient to reduce intracranial tumor growth [60], possibly due to poor gene transfer in established tumor tissue. Interestingly, adenoviral transduction of p53 into wild type p53-bearing glioma cells has also shown marginal [59–61] to robust [62] inhibition of cell proliferation and induction of apoptosis, indicating that these effects are not solely dependent on the restoration of a functional copy of p53.

Limitations of tumor suppressor-based gene therapy have included poor gene transfer as mentioned, lack of bystander effect, and potential resistance arising from the inherent genetic heterogeneity within GBM. However, transduction of tumor suppressor genes such as p53 may present an excellent opportunity for combinatorial therapy since they could re-sensitize the cells to radiation and chemotherapy [59,61,63,64] or reduce immune evasion when combined with immune-boosting strategies [65]. Successful pre-clinical results with p53-restoration led to a phase I clinical trial of Ad5CMV-p53 (INGN 201) for recurrent malignant glioma, involving injection of the virus pre- and post-resection. Exogenous p53 protein was found in the nuclei of tumor cells in all patients treated with this strategy, although transduced cells were found only within a short distance from the injection site. Adverse events were minimal and the trial tested doses up to 3×10^{12} plaque-forming units (p.f.u.) without reaching maximum tolerable dose [66].

$p16^{INK4A}$ is another major tumor suppressor that causes cell cycle arrest at the G1-S transition point by maintaining hypo-phosphorylated status in the Retinoblastoma protein (Rb) [59,60]. Adenoviral-mediated restoration of p16 in GBM cells induced, as expected, tumor cell cycle arrest in G1-S phase [67,68]. Surprisingly, this overexpression of p16 also caused an unexpected reduction in GBM cell invasion, resulting from decreased activity of matrix metalloprotease 2. This important anti-invasive effect is remarkable since it was not observed after restoration of the tumor suppressor p21/WAF, which prevents G1-S transition by a molecular mechanism similar to p16 [67].

The third major tumor suppressor regulating glioma growth and invasion is the Phosphatase and Tensin Homologue (PTEN), which is lost, mutated or inactivated in 40%–50% of all gliomas (~25% of primary GBMs) [58], resulting in high levels of dys-regulated PI3K activity and downstream signaling [69]. Adenoviral re-expression of PTEN in GBM cells inhibited Akt kinase activity, leading to tumor cell apoptosis [70]. Infection with this virus was also shown to decrease metalloprotease expression and glioma cell invasion *in vitro* [71]. When tested in GBM-bearing mice, adenoviral restoration of PTEN has shown important effects on the tumor microenvironment, inducing an anti-angiogenic response even in presence of pro-angiogenic stimuli such as loss of p53 or presence of constitutive EGFR activity [72].

Another important example of viral-delivered tumor suppressor strategy has been demonstrated with p27, an inhibitor of Rb phosphorylation that arrests the cell cycle in G1. p27 levels are regulated by complex feedback loops involving phosphorylation of this protein in Thr^{187} and further proteasomal degradation [73]. p27 activity was restored in GBM cell lines and GBM-derived primary cells using adenovirus to carry either wild type (Ad-p27wt) or a degradation-resistant Thr^{187}-mutant (Ad-p27mt, $\text{Thr}^{187}\text{Ala}$). In all cases, recovery of functional p27 promoted Rb dephosphorylation, apoptosis, and suppression of tumor growth [68]. Interestingly, while p27wt arrested the cell cycle in G1-S transition as expected, p27mt did so at the G2-M checkpoint by undefined mechanisms that were not observed in other cell types. Additional studies using adenoviral-restored p27 have demonstrated that this tumor suppressor reduces GBM growth *in vivo* as well as local invasion and tumor-induced neo-angiogenesis, with these effects being caused by a cytoskeletal anti-migratory effect of p27 both in GBM cells and tumor-associated endothelial cells [74].

2.3. Viral Delivery of Immunomodulatory Genes

The CNS is relatively isolated from systemic immune responses and is therefore difficult to induce the immune system to mount an effective local anti-tumor response against gliomas [42]. This difficulty is increased by the ability of glioma cells to suppress and effectively evade cellular immune responses [9]. In order to promote effective immunotherapy against glioma, viruses have been engineered for targeted delivery and expression of cytokines that activate and recruit immune effectors to the tumor.

An excellent example of this strategy was the early use of a replication-deficient adenovirus carrying the gene for the potent immune-boosting cytokine interferon beta (IFN-beta) under control of the CMV promoter [75,76]. Using this vector, IFN-beta was expressed in pre-established, subcutaneous gliomas in nude mice, resulting in enhanced immune cellular response against the tumor (lymphocyte infiltration), tumor regression, and significantly prolonged animal survival [76]. This strategy reached a phase I clinical trial for recurrent malignant glioma, where an IFN-beta-expressing, non-replicating adenoviral vector was stereotactically injected in the tumor before surgical resection [77]. The trial demonstrated that the virus inoculation was safe and well tolerated, while analysis of the resected tumors demonstrated dose-dependent induction of local inflammation and tumor necrosis.

Using recombinant parvoviruses, another immunomodulatory strategy was attempted by simultaneous delivery of IFN-gamma-inducible protein 10 (CXCL10) and TNF-alpha in a syngeneic mouse model of GBM [78]. Results showed synergistic activity of both vectors and complete regression of tumors generated from cells that had been transduced with both cytokines before implantation. Multiple mechanisms were proposed to contribute to this synergy, including CXCL10-mediated recruitment of activated T and NK lymphocytes to the tumor, inhibition of tumor angiogenesis by CXCL10, and TNF-alpha-mediated tumor necrosis and maturation of dendritic cells. Despite these exciting results, the effect of the viruses in naïve pre-established tumors was marginal, resulting in delayed tumor growth but no regression [78].

Viral-mediated delivery of other interleukins has not been as extensively exploited in GBM as in other cancers, but work in this direction has definitely shown therapeutically relevant results. Non-replicating adenoviral-associated virus (AAV) and replicating HSV have been employed to deliver IL-12 in experimental models of GBM [79–81], resulting in local immune mechanisms such as increased IFN-gamma expression, microglial activation, and recruitment of T and NK lymphocytes, with a significant antitumor effect.

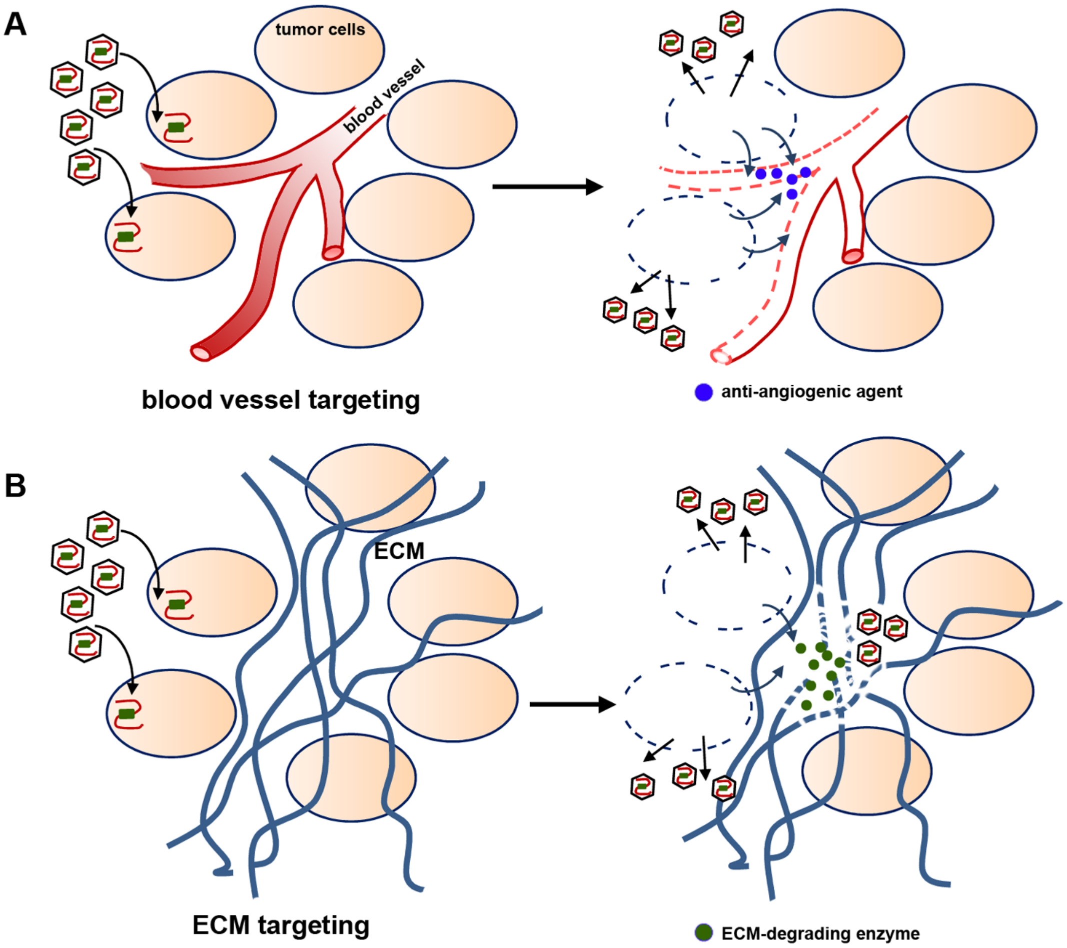
2.4. Viral Delivery of Genes That Modify the Tumor Stroma

The gene-delivery strategies described in the previous sections (as well as viral-mediated oncolysis, in the following section) target specifically the tumor cells for immediate cell death. Additional effects such as reduced tumor vascularization and invasion may be observed (and welcomed) but are not usually part of the design rationale. However, viruses can also be engineered to deliver genes that specifically affect the tumor microenvironment. Two clear examples of this strategy are viruses carrying anti-angiogenic genes or genes that remodel the tumor extracellular matrix (as illustrated in Figure 2).

Initial attempts to specifically inhibit glioma angiogenesis with gene therapy involved the intratumoral injection of retrovirus and AAV carrying the antiangiogenic factor angiostatin [82,83]. A subsequent study, using systemic instead of local delivery, followed a similar approach with adenoviral-delivered endostatin [84]. In all cases tumor vascularization was significantly inhibited and tumor growth was reduced more effectively than with the parental viruses. A more recent approach has combined anti-angiogenesis with viral oncolysis using conditionally-replicating oncolytic HSV (see details of this virus in the following section). Two oncolytic HSVs were engineered to express the anti-angiogenic protein vasculostatin under control of the promoter for the early viral gene IE4/5 [85,86]. Secretion of vasculostatin was detected a few hours after infection of glioma cells and results *in vivo* with both viruses showed remarkable reduction in microvessel density, tumor perfusion, and overall tumor progression. Coupled with their oncolytic ability, these antiangiogenic viruses offered significantly better antitumor efficacy when compared with their parental HSV strain.

Delivery of genes that can remodel the tumor extracellular matrix (ECM) has been rarely attempted, despite the fact that this matrix is the most immediate physical barrier to viral dispersion and a significant factor that limits viral effects to short distances from the inoculation site [87]. Indeed, pre-injection of proteases that can degrade tumor ECM proteins enhances subsequent viral spread and infection [88]. Following this rationale, a conditionally-replicating oncolytic HSV was engineered to express the bacterial enzyme chondroitinase ABC-I, which degrades major components of the glioma ECM such as hyaluronan and chondroitin sulfate proteoglycans [89]. The resulting virus demonstrated increased efficacy against the tumor compared to parental HSV and, as expected, dispersed farther away from the sites of inoculation. Local degradation of ECM was also demonstrated, which, importantly, did not enhance the invasive ability of the remaining tumor cells.

Figure 2. Strategies targeting the GBM microenvironment. To enhance viral oncolysis conditionally-replicating oncolytic viruses may also carry genes that modify the tumor microenvironment. (**A**) Anti-angiogenic strategies: viruses carry anti-angiogenic factors that reduce vascular support of the spared tumor not reached by oncolysis; (**B**) Anti-ECM strategies: viruses carry enzymes that degrade ECM components, increasing dispersion of viral particles and oncolytic efficacy.



2.5. Replication-Competent Oncolytic Viruses

Oncolytic virotherapy (OV) can be considered a gene therapy approach where the whole virus becomes the genetic payload. OV utilizes replication-competent viruses that can infect and lyse tumor cells, with or without concomitant gene transfer [9,90,91]. OV strategies exploit two major properties of the viruses used as vehicles/lytic agents: Tropism towards cells of neural lineage/brain tumor cells, and the ability to specifically replicate in tumor cells with altered signaling pathways while sparing normal cells [90,92–94]. Oncolytic HSV, conditionally-replicating adenovirus (CRAd), reovirus, poliovirus, engineered retroviruses, Newcastle Disease virus and measles virus have been evaluated for OV therapy of GBM [95].

HSV is an enveloped DNA virus with wide tropism that replicates in dividing and non-dividing cells [9,91]. This virus establishes latent infection in post-mitotic neurons and is especially suitable for therapeutic strategies in the CNS. To reduce neurovirulence, engineered versions have been attenuated through deletions in genes necessary for viral replication in normal cells, such as the ribonucleotide reductase ICP6 (*UL39*) and the protein synthesis promoting factor ICP34.5 (*γ34.5*) [96]. Two recombinant oncolytic HSV have reached clinical trials for recurrent high-grade glioma: The virus HSV G207 (with an inactivating insertion in ICP6 and deletions in both copies of γ34.5) was tested in a phase I clinical trial where no dose-limiting toxicity was observed up to the maximum tested dose of 3×10^9 p.f.u. The study reported eight patients (out of 21) with radiographic/histologic response to the treatment and two long-term survivors [97]. A further phase Ib trial demonstrated the safety of multiple dose delivery of the same virus, including inoculation both in the pre-resected tumor and the post-resection cavity [98]. A similar virus, HSV1716 (deleted in both copies of γ34.5) reached phase II trial for intratumoral delivery. HSV1716 treatment showed positive response in three out of twelve patients and did not cause toxicity even when patients developed antibodies against the virus [99,100]. A disadvantage of these models is the required deletion of viral genes (γ34.5), which attenuates viral replication and limits the virus efficacy [101]. A new generation of oncolytic HSVs is therefore being studied in pre-clinical models and prepared for clinical trials. A major modification that enhances the efficacy of these viruses is the conditional re-expression of viral genes under promoters that are overexpressed in brain tumor cells compared to normal cells, such as Nestin, GFAP, or survivin [102–104].

CRAds with deleted viral genes have been employed as alternative to oncolytic HSV. Adenoviruses are non-enveloped DNA viruses that infect proliferating and quiescent cells and can integrate, with low frequency, into a defined region of the host genome [9,91]. Two extensively studied CRAds in glioma are ONYX-015 and Ad5Delta24, both of which target cells with dys-regulated signaling pathways. ONYX-015 has a deletion in the gene coding for the viral protein E1B that binds and inactivates p53. Due to this deficiency, the virus was originally expected to replicate selectively in p53-deficient cells (*i.e.*, tumor cells) [105], although it was later shown that its oncolytic activity in gliomas was independent of p53 status and even increased in wild-type p53 glioma xenografts [106]. ONYX-015 was tested in a phase I trial for recurrent malignant gliomas, being injected only in the post-resection cavity [107]. Although the study could not demonstrate a significant antitumor efficacy, it showed absence of serious adverse effects and good tolerance to the virus, without reaching maximum tolerable dose even at 10^{10} p.f.u.

Similarly to ONYX-015, Ad5Delta24 has a genetic deletion for the viral protein E1A, which inactivates the Rb tumor suppressor. Ad5Delta24 therefore replicates in glioma cells with a deficient Rb pathway, causing significant growth inhibition of xenografted tumors in mice [108]. This virus was further engineered by introducing the integrin-binding RGD motif in the knob domain of the viral fiber protein (Ad5Delta24-RGD) [109], which enhanced the specific targeting of tumor cells and increased oncolytic efficacy against gliomas [110]. Ad5Delta24-RGD is currently being tested in a phase I clinical trial for recurrent malignant glioma (NIH trial NCT00805376, Table 1) [111].

A novel CRAd that has been recently developed replaces the strategy of viral gene deletion by using instead the promoter of a gene highly expressed in glioma cells (survivin) to drive the expression of E1A [112]. This virus (CRAd-survivin-pK7) has in addition a poly-lysine sequence added to the fiber knob, which enhances the binding to cell surface proteoglycans and increases viral entry and anti-tumoral efficacy against differentiated as well as stem-like glioma cells [112,113].

In addition to the extensively used HSV and CRAds, other viruses have been engineered for OV of glioma, including measles, vaccinia, rhabdoviruses and polioviruses [114,115]. Of these, the measles paramyxovirus (MV, attenuated Edmonston strain) was the earliest to be considered as a potential oncolytic agent since initial reports in the 1970s suggested that measles infection was associated with regression of lymphoma and leukemia [116,117]. MV targets cells that express the membrane receptors CD46 and Signaling Lymphocyte Activating Molecule (SLAM), inducing the formation of multinucleated syncytia followed by apoptosis. MV variants have been engineered to express IL-13 [118] or a single-chain antibody against the vIII deletion variant of EGFR [119], therefore re-targeting the viruses against proteins highly expressed on the surface of glioma cells and increasing their oncolytic efficacy. Additional engineering of MV to express the circulating carcinogenic embryonic antigen (MV-CEA) has been used to monitor the course and maintenance of MV infection [120]. Toxicological data in macaques has shown absence of neurotoxicity of MV-CEA [121] and supported an ongoing phase I clinical trial (NIH trial NCT00390299, Table 1). Recent data has also shown that MV is highly effective against glioma-derived stem-like cells [122], which makes it an attractive approach against this highly resistant population of tumor cells.

2.6. Advantages and Challenges of Viral-Based Gene Therapy

Having evolved for horizontal gene transfer, viruses are the most efficient carrier system to deliver genes to tumor cells. Additional modifications described in this review, such as re-targeting and conditional replication have considerably improved the specificity and efficacy of viral vectors, many of which have reached clinical trials for GBM. Moreover, compared to other particle-like carriers (see Sections 4.1–4.3 on nanotechnology), viruses induce robust bystander cytotoxic effect, attract cellular immune response towards the infected cells, and can directly kill infected glioma cells by cell lysis. These features absent in other vehicles for gene delivery have made viruses one of the most valuable tools for gene therapy of GBM.

On the other hand, viral carriers and oncolytic viruses still face considerable challenges for successful long-term therapeutic effects. A major difficulty is the limited spread and persistence of the virus in the tumor tissue, caused by factors such as low efficiency of initial infection, rapid clearance of the viral particles by innate immune cells, and physical barriers that limit particle dispersion [123]. Some of these challenges are being actively addressed through strategies involving viral engineering and combination with other antitumor agents. Major developments in the field include improved re-targeting towards GBM-specific receptors [124], combination with drugs that reduce the immune response to the virus [125], "cloaking" of the virus inside carrier cells (described in the following section), enhanced infusion of viral particles via convection-enhanced delivery, and engineering of viruses to express genes that facilitate their physical dispersion [89], among others. As current limitations are overcome, viral-based approaches (alone or combined with conventional therapy) will remain a major choice for gene therapy of GBM.

3. Stem Cell-Based Gene Therapy of GBM

Together with viruses, cells have been used to deliver genetic material to brain tumors for the past twenty years. Indeed, retroviruses themselves were the first genetic payload delivered by cells injected into the tumor stroma [18,19]. Examples of successful carriers have included fibroblasts and HEK293 cells, both used to deliver replicating viruses, suicide genes (HSV-TK), and anti-angiogenic factors in gliomas [126,127]. A major limitation of these cell types has been their lack of migratory ability inside the tumor, a deficiency that was considered a major cause of therapeutic failure in clinical trials of viruses and suicide genes delivered by cells [37]. Currently, the most widely employed cellular carriers are stem cells (SCs) of neural, mesenchymal or embryonic origin. While SCs have been studied for only half as long as viral carriers and only recently reached the clinical stage, they have proved one of the most attractive vehicles to combine gene therapy with virotherapy and conventional therapies. The importance of SCs is underscored by a fundamental property absent in other delivery vehicles: their ability to migrate towards the tumor cells even when injected peripherally [128,129]. This critical feature allows them, in principle, to reach the disseminated tumor cells that are characteristic of GBM [128]. As expected, SCs have already been extensively tested as vehicles for most of the approaches described in the previous sections: delivery of suicide genes, oncolytic viral particles, anti-angiogenic factors, and immune-boosting cytokines, among others.

3.1. Neural Stem Cells

Neural Stem Cells (NSCs) are multipotent progenitors of the neural lineage with indefinite self-renewal and ability to differentiate into neurons or glial cells [130]. They are not only highly adapted to the neural environment and architecture [131] but also share many properties (such as cell motility mechanisms [119]) with the elusive glioma stem-like cells. Engineered NSCs were first used against gliomas in 2000 to deliver the cytokine IL-4, therefore improving the immune response against the tumor [132]. Subsequently, they have been largely employed in two major approaches for antitumor gene therapy: as infected carriers of oncolytic viruses or as engineered cells expressing therapeutic genes.

Using NSCs as carriers of oncolytic viruses -which lyse the carrier cell and infect glioma cells- has become an interesting approach with multiple possible advantages over the inoculation of viral particles: migratory NSCs may deliver the viruses at further distances within the tumor compared to virus alone; they can protect the viruses from the host immunosurveillance; and their own lysis removes them from the host after therapy [133]. Following this rationale, Herrlinger and colleagues were the first to show the feasibility of using NSCs to carry conditionally-replicating HSV into pre-implanted cerebral gliomas [134]. Similar studies were further pursued with CRAds [135,136], demonstrating that pre-loading the virus inside NSCs highly enhanced the reach of the virus within the tumor as well as its oncolytic efficacy.

Engineering NSCs to deliver transgenes rather than viruses into the tumor mass has been a more common approach, employed with suicide genes and cytokines. Both CDA and HSV-TK (described in Section 2.1) have been tested in NSC therapy of GBM (e.g., [128,137,138]) and in all cases the use of NSCs has compared favorably against non-migratory cell carriers. NSCs have also been used to carry CDA and IFN-beta cDNAs together, boosting the bystander cytotoxicity with immune response against the tumor and resulting in better antitumor response compared to CDA alone [139]. CDA-carrying NSCs are currently being tested in the first clinical study of feasibility against recurrent high-grade gliomas (NIH trial NCT01172964, Table 1).

A second group of widely tested transgenes have been immune-boosting interleukins such as IL-4, IL-12 and IL-23 [132,140,141]. These studies led to two important conclusions: First, NSC-mediated sustained delivery of interleukins was found more efficient than viral-based delivery *in vivo* [132] and resulted in improved animal survival. Second, the strong cellular immune response against the tumor (at least in the case of IL-23) resulted in long-term surviving animals resistant to tumor re-challenge, suggesting the possibility of using SC approaches to trigger long-standing antitumor immunity.

One important cytokine that has been delivered by NSCs in gliomas is the Tumor Necrosis Factor-Related Apoptosis-Inducing Ligand (TRAIL), which is capable of inducing apoptosis in tumor cells with little effect on normal cells. This effect is potentiated in S-TRAIL, a secreted chimeric protein combining the extracellular domains of TRAIL and Flt3L (a ligand for Flt3 tyrosine kinase receptor) [142]. NSCs carrying S-TRAIL were shown to migrate actively towards glioma cells, causing bystander cytotoxicity in the tumor and significantly reducing the burden of pre-implanted gliomas [143]. Moreover, combination of NSCs-S-TRAIL with adjuvant chemotherapy showed strong potentiation of the antitumor effects of temozolomide, PI-103 (a PI3K/mTOR inhibitor), and bortezomib (a proteasome inhibitor) [144–146].

A third major group of genes delivered by NSCs in glioma models includes anti-angiogenic factors such as endostatin [147], thrombospondin-1 [148], and the angiostatic factor PEX [149]. Although these factors can impair the migration of normal and tumor cells, all the studies reported that active migration of the engineered NSCs towards tumor cells was unaffected. NSCs distributed extensively in the tumor stroma, without concomitant proliferation or differentiation, and caused significant reduction of tumor growth and microvascular density.

3.2. Mesenchymal Stem Cells

Mesenchymal stem cells (MSCs) are non-hematopoietic, adult multipotent stem cells that can be isolated from multiple sources (such as bone marrow, adipose and muscle tissue, or peripheral blood), and expanded with relative ease *in vitro* [150]. These cells migrate to sites of injury and inflammation and are involved in tissue repair. Since tumors behave as non-healing, expanding wounds, they strongly attract MSCs [151]. Based on this strong homing ability, MSCs have been employed in multiple strategies against tumors, including gliomas [152].

As with NSCs, initial studies of MSCs against experimental gliomas involved the overexpression of a therapeutic cytokine (IL-2), resulting in augmented antitumor effect (compared to MSCs alone) and prolongation of animal survival [153]. Due to their easy availability for autologous transplantation, MSCs have become a very attractive alternative to NSCs towards clinical studies, and have already been tested to deliver cytokines, suicide genes, antibody chains and viral particles to gliomas.

Suicide genes originally tested in NSCs have been evaluated in MSCs as well, including HSV-TK [144], CDA [154], and HSV-TK together with connexin-43 to enhance bystander cytotoxicity [145]. MSCs have also been used as vehicle for the suicide gene carboxylesterase [155]. A combination of MSC-delivered carboxylesterase with the prodrug CPT-11 (Irinotecan) was tested in experimental brainstem gliomas and showed effective conversion of the prodrug but only modest improvement in animal survival.

In addition to MSC-delivered IL-2, more recent animal studies have used transduced MSCs to deliver IL-7 [156], IL-12 [157], or IL-18 [158] in glioma xenografts. Strong cellular immune response was observed following infiltration of the MSCs in the tumor stroma, and resistance of the surviving animals to tumor re-challenge was observed in the studies using IL-12 and IL-18. MSCs have also been used to deliver TRAIL constructs, including a shortened form of secretable TRAIL [159] as well as the recombinant TRAIL-Flt3L fusion (S-TRAIL) [160]. Newer studies have combined MSC-delivered TRAIL with radiotherapy [161] and with the anti-inflammatory compound MK886 [162], demonstrating in both cases synergism compared to MSC-TRAIL alone.

Finally, the use of MSCs as carriers of conditionally-replicating oncolytic viruses such as CRAds has shown that MSCs are capable of suppressing the humoral immune response against the virus, resulting in increased persistence of the viral particles [163]. However, MSCs were found less efficient than NSCs to deliver the same CRAd in orthotopic xenografts and had lesser impact on animal survival [164]. Nevertheless, considerable interest remains in the use of MSCs to deliver those oncolytic viruses that have reached clinical trials, such as Ad5Delta24-RGD described in Section 2.5. MSCs carrying Ad5Delta24-RGD and injected in the peripheral circulation have been shown to reach intracranial gliomas and inhibit tumor growth [165]. Current trends suggest that, if clinical trials with the viruses are promising, the use of SCs to carry them into the tumor could be the next step to improve their therapeutic efficacy.

3.3. Embryonic Stem Cells

Embryonic stem cells (ESCs) are pluripotent cells that form the inner cell mass of the blastocyst during gestation and have unlimited proliferative capacity [166]. Difficulties in obtaining and culturing these cells—both technical and derived from current legislation on SCs—have made their use much more limited than NSC or MSC counterparts, with experiments largely focused on cytokine delivery to gliomas.

The first experiments with mouse ESCs involved engineering to express TRAIL, followed by differentiation into astrocytes using defined culture conditions. ESC-derived astrocytes enhanced apoptosis of co-cultured glioma cells compared to TRAIL alone [167]. Further experiments demonstrated that injection of ESC-derived astrocytes carrying TRAIL could induce severe necrosis in xenografted tumors [168], but no survival studies were pursued. Additional *in vitro* studies using ESC-derived astrocytes demonstrated that delivery of a different cytokine (IL-24) also increased apoptosis of co-cultured glioma cells and potentiated the effects of radiation and temozolomide [169].

Human ESC lines have more recently been used to derive NSCs and MSCs, which have been subsequently transduced to deliver the suicide gene HSV-TK [170,171]. The goal of these experiments was to show the successful conversion of ESCs into cells known to have strong tropism for gliomas, and in addition these studies demonstrated successful targeting of intracranial tumor burden and extension of animal survival.

3.4. Advantages and Challenges of Cell-Based Gene Therapy

The use of SCs as gene-delivery vehicles is supported by two unmatched advantages when compared to passive methods of gene delivery: (a) migratory ability that allows them to infiltrate the tumor mass, reaching poorly vascularized areas and the remote borders of the tumor; and (b) strong tropism that attracts them towards glioma cells even when injected peripherally, coupled with ability to cross the blood brain barrier. These two features of SCs, added to the possibility of performing extensive genetic engineering to convert them in carriers of multiple transgenes or whole viral vectors, make them a versatile tool that can be combined with conventional therapy and additional molecular therapy to deliver a large, complex payload inside the tumor.

However, despite their ability to infiltrate gliomas, SCs are essentially neutral and do not have an effect on the tumor unless engineered as gene-delivery vehicles. Since the transgenes are expressed in SCs immediately after transduction (in contrast to viral-carried genes, which are expressed only after infection of the target cells), a first and considerable technical challenge is to ensure that the SCs will survive for as long as it takes to impact the tumor cells, without dying first due to effects of suicide genes or oncolytic viruses [172]. Rapid and efficient delivery to the tumor is therefore a critical factor when SCs are introduced peripherally. Intravenous injection has been the most common route for peripheral introduction of SCs but its efficiency is limited, with less than 2% of the inoculated cells colonizing the tumor [173]. A recent alternative has used intranasal inoculation of NSCs, with a delivery efficiency estimated to be as high as 24% [174]. Additional challenges stem from the choice of SCs in terms of convenience, permanence in the tumor, and therapeutic efficacy. For example, while MSCs are easiest to obtain for autologous therapy, there is active discussion about their relative efficacy compared to NSCs for different gene-therapy strategies [164]. ESCs present, in addition, ethical and regulatory issues for collection and will likely be replaced by induced pluripotent SCs in the future.

A final and considerable factor that must be addressed with SCs is their safety when introduced in the highly aggressive, cytokine- and growth factor-rich environment of the tumor. To this day studies have shown that none of the different types of SCs employed in animal models suffered neoplastic transformation. However, previous studies have demonstrated that normal neural progenitor cells can contribute significantly to the heterogeneous total mass of PDGF-induced malignant gliomas [175]. Therefore, a desirable feature in future SC-based approaches would be the possibility of selectively eliminating the SCs (e.g., using an inducible suicide gene) after they have reached their therapeutic endpoint.

Overall, SC-based gene therapy of GBM offers enormous promise and, considering that SCs have become the choice carrier in other neuropathologies, is likely to become the fundamental component of future combinatorial strategies using gene delivery, molecular-targeting therapy and conventional chemoradiotherapy.

4. Nanotechnology-Based Gene Therapy of GBM

The use of nanotechnology, *i.e.*, manipulation of sub-micron-sized materials, to target genetic material into tumor cells is a relatively novel strategy that remains largely experimental. The only nanocarriers that have reached the clinical stage in glioma have been liposomes [9,176], which have long been used as carriers for small molecules in glioma and other cancers. Most nanotechnological approaches for gene therapy have focused in optimizing the DNA-carrying vehicles for effective targeting of tumor cells [177,178], testing many of the same candidate genes used with viruses and SCs. Vehicles that have been recently tested in pre-clinical models include novel formulations of cationic liposomes, nanoparticles, and dendrimers, among others.

4.1. Liposomes

Liposomes are artificial, lipid-based microvesicles usually employed to deliver drugs, peptides and proteins into cells. However, chemical engineering of the lipids also permits the formation of stable DNA-lipid associations that can be exploited to use liposomes as a gene-delivery vehicle. Following this concept, a liposomal vector was devised in the early 2000s to carry a plasmid coding for HSV-TK, which was given to patients with recurrent GBM in a phase I/II trial via intratumoral infusion [179], followed by administration of the prodrug ganciclovir during 14 days. Radiographic response was observed in most patients in this clinical trial, ranging from focal effects to 50% reduction in tumor volume, without major adverse events.

Cationic liposomes have also been used to transfer cytokine genes into glioma cells. A phase I/early phase II clinical trial demonstrated the safety and efficacy of this approach to deliver a plasmid coding for IFN-beta [180] in patients with recurrent malignant gliomas, following resection of the tumor. IFN-beta protein was found in the accumulated fluid in the post-surgical cavity in three out of five patients, for periods of up to ten days after injection. All patients showed positive radiographic response immediately after treatment, although the tumors eventually progressed. A more complex liposome was recently devised to carry both a therapeutic gene (TRAIL) and a cytotoxic drug (paclitaxel), combined with re-targeting via addition of a peptide (angiopep2) that facilitates blood brain barrier crossing. This liposome preparation effectively delivered TRAIL to glioma cells *in vitro* even when protected behind a barrier of normal epithelial cells [181]. More importantly, peripherally delivered angiopep2-liposomes reached intracranially xenografted gliomas in mice, causing local apoptosis in the tumor and extending animal survival.

Other liposomal preparations have modified not only the surface of the vesicle but also its carrier core. For example, magnetite-core cationic liposomes respond to alternating magnetic fields generating heat. This effect can be used to activate a heat-shock sensitive promoter in the DNA carried by the liposome, thus regulating expression of the therapeutic gene [182]. Using this strategy, Ito and colleagues injected paramagnetic liposomes carrying TNF-alpha under control of the heat-inducible promoter *gadd153* into subcutaneous gliomas implanted in nude mice [183]. Their results showed heat-regulated expression of TNF-alpha protein in the tumor and subsequent retardation of glioma growth.

A recent and ingenious liposomal design used air-cored liposomes, subsequently loaded with siRNA against the anti-apoptotic protein sirtuin 2 [184]. The liposomes were injected in subcutaneously-implanted gliomas in nude mice, followed by exposure to brief low-frequency ultrasound. This induced cavitation (bursting) of air bubbles in the liposomal core, damaging neighboring tumor cells and enhancing the delivery of therapeutic siRNA. The results demonstrated effective decrease in tumor volume and prolonged animal survival when compared to liposomes lacking siRNA or absence of ultrasound treatment.

4.2. Polymers

Cationic polymers are macromolecules that spontaneously bind DNA via electrostatic interactions. This unique property has been used commercially for cell transfection. They offer advantages such as small size and flexible chemistry that allows extensive modifications to improve biodistribution and tumor targeting.

A typical example of a linear polymer used to deliver plasmids or oligonucleotides is polyethylenimine (PEI). This polymer binds DNA strongly and has high transfection efficiency, forming small particles that enter the cells via endocytosis [185]. However, in absence of further modifications it has high cellular toxicity and cannot reach intracranial tumors when injected peripherally. Chemical engineering of PEI by addition of functional groups such as poly-ethileneglycol (PEG) or beta-cyclodextrin has proven sufficient to improve PEI permanence in circulation and in the tumor stroma [186]. PEI polymers modified by addition of myristic acid were able to cross the blood brain barrier, delivering a TRAIL-coding plasmid to intracranially implanted gliomas and increasing survival in tumor-bearing mice [187]. Similarly, PEGylated PEI was re-targeted towards glioma cells by chemical addition of an RGD-containing peptide [188]. This polymer (RGD-PEG-PEI) was injected intravenously and able to deliver TRAIL cDNA *in situ* in an intracranial glioma model, increasing animal survival.

In addition to the chemically simpler linear polymers, novel efforts have focused on using repeatedly branched polymers, known as *dendrimers*, for gene delivery. These molecules offer several advantages (such as high surface/volume ratio for DNA binding and well-known chemical behavior) [189] that have made them attractive synthetic nanocarriers for gene therapy. A commonly used dendrimer is a hyperbranched polymer of poly-amidoamine (PAMAM) characterized by biocompatibility, controlled biodegradation, low toxicity, and good accumulation in tumors with leaky vasculature [190]. A modified version of PAMAM was conjugated with nanoparticle carriers (see next section) and a viral Tat-peptide to facilitate cell membrane crossing [191]. This complex polymer (np-PAMAM-Tat) was used to deliver anti-EGFR shRNA to subcutaneously-implanted gliomas, inhibiting EGFR/Akt signaling and slowing tumor growth. Another modified version of PAMAM (Arg-PAMAM) has recently been used to deliver IFN-beta cDNA in intracranial glioma xenografts, causing selective tumor cell apoptosis and overall tumor shrinkage [192].

4.3. Nanoparticles

As their name indicates, these are nanometer-sized particles that, depending on their size (usually ranging from 20 to 50 nm in diameter) may spontaneously cross capillary walls and be endocytosed by cells. They have a rigid polymer core and a multi-functionalized surface that has been engineered to enhance DNA binding, particle diffusion, and cell-membrane crossing [193]. The core of the nanoparticles can also be modified with fluorochromes or metallic iron to facilitate *in vivo* imaging or to make them responsive to magnetic stimuli [194,195].

Super-paramagnetic iron oxide nanoparticles (SPIONs) have been used for delivery of therapeutic agents to the CNS while being at the same time tracked via magnetic resonance imaging [196]. In a recent study, SPIONs were covalently bound to the capsid of an RGD-targeted adenovirus (replication-deficient Ad5/3-RGD) and used to track the delivery of the virus [197]. Results demonstrated that viral infectivity of glioma cells was not affected by the SPIONs and that the virus could be tracked in a large (porcine) brain, but no studies were performed with gliomas *in vivo*.

More complex version of SPIONs have used magnetic nanoparticles simultaneously functionalized on their surface with PEG (to bypass the blood brain barrier), chitosan (to improve half-life in circulation), PEI (to adsorb the therapeutic DNA) and chlorotoxin, a re-targeting peptide against glioma cells [198,199]. These multifunctional SPIONs have been used to demonstrate proof-of-principle delivery of siRNA and marker cDNA (eGFP) to glioma cells and subcutaneously-implanted gliomas.

4.4. Advantages and Challenges of Nanocarrier-Based Gene Therapy

Being completely synthetic, nanocarriers offer strong advantages as vehicles for gene therapy of GBM: size, structure and chemical composition of the carrier can be accurately tailored to enhance biodistribution, low toxicity, and optimal cell penetrance. The size of the DNA that can be carried is not as limited as in biological carriers, does not require prior genetic engineering of the carrier, and allows in principle any combination of plasmids and oligonucleotides as desired. The surface of nanocarriers is based on well-studied polymers that are biocompatible, biodegradable, and do not induce immune responses against the vehicle, as it is common when using viral carriers. As an additional advantage, the core of the carrier can also be optimized to track it using fluorescence or magnetic resonance, a property that would require additional modifications in viral or cell-based carriers.

On the negative side, the major disadvantage of nanocarriers is that they are completely passive vehicles for gene delivery and their efficacy depends on the physical and chemical properties of the materials used to build them. Nanocarriers do not migrate actively and do not have neurotropism or even cell-specific tropism, therefore the tumor specificity must be "built from scratch" and optimized more exhaustively than with SCs or viruses. Their passivity as delivery vehicle also means that their distribution in the tumor will be much more limited than that observed with SCs and even with viral carriers, which can disperse when they lyse infected tumor cells. Finally, biological effects of novel particles and polymers (and their fragments) have not yet been analyzed in long-term studies and the overall efficacy of nanocarriers has not been exhaustively compared against other modalities for gene delivery. The future of these experimental vehicles for gene therapy of GBM will depend on these efficacy studies, using nanocarriers with improved cell type-specificity delivered via convection-enhanced delivery and combined with conventional therapies.

5. Conclusions

Gene therapy represents today one of the more flexible and robust strategies for adjuvant therapy of GBM. As detailed in this review, gene-delivery approaches can be manipulated at multiple levels including choice of delivery vehicle, chemical or genetic engineering of the carrier, and selection of molecular targets, among others. This wide range of manipulations can be extensively exploited to optimize biodistribution, persistence, specificity, and targeting effects of the therapeutic agents, arguably to a much further extension that what can be achieved with improvements in conventional chemoradiotherapy. Optimized gene therapy approaches have repeatedly reached the clinical stage: A brief query of clinical trials listed by the Journal of Gene Medicine Clinical Trial website [11] and trials registered with the US National Institutes of Health [10] reveals, at the time of this writing, over a dozen active studies to deliver genetic material in malignant gliomas (Table 1). This underscores the extent to which gene therapy has become one of the most important approaches in molecularly-targeted therapy for GBM, and indicates that future developments in this field will continue improving the therapeutic options for this devastating disease.

Acknowledgments

The authors wish to acknowledge fruitful discussions with all members of the Chiocca and Viapiano laboratories in the Department of Neurosurgery, Brigham and Women's Hospital. This work was supported by grant 5R01CA152065 to MSV.

Conflicts of Interest

The authors declare no conflict of interest. E.A.C. is involved in two listed active trials using adenoviral-delivered HSV-TK.

References

1. Wen, P.Y.; Kesari, S. Malignant gliomas in adults. *N. Engl. J. Med.* **2008**, *359*, 492–507.
2. Dolecek, T.A.; Propp, J.M.; Stroup, N.E.; Kruchko, C. Cbtrus statistical report: Primary brain and central nervous system tumors diagnosed in the united states in 2005–2009. *Neuro-oncology* **2012**, *14*, v1–v49.
3. Schwartzbaum, J.A.; Fisher, J.L.; Aldape, K.D.; Wrensch, M. Epidemiology and molecular pathology of glioma. *Nat. Clin. Pract. Neurol.* **2006**, *2*, 494–503.
4. Louis, D.N. Molecular pathology of malignant gliomas. *Ann. Rev. Pathol. Mech. Dis.* **2006**, *1*, 97–117.
5. Nduom, E.K.; Hadjipanayis, C.G.; van Meir, E.G. Glioblastoma cancer stem-like cells: Implications for pathogenesis and treatment. *Cancer J.* **2012**, *18*, 100–106.
6. Keeler, C.E. Gene therapy. *J. Hered.* **1947**, *38*, 294–298.
7. Friedmann, T.; Roblin, R. Gene therapy for human genetic disease? *Science* **1972**, *175*, 949–955.
8. Sheridan, C. Gene therapy finds its niche. *Nat. Biotechnol.* **2011**, *29*, 121–128.
9. Natsume, A.; Yoshida, J. Gene therapy for high-grade glioma: Current approaches and future directions. *Cell Adhes. Migr.* **2008**, *2*, 186–191.
10. US National Institutes of Health. Index of clinical trials. Available online: http://clinicaltrials.gov (accessed on 3 October 2013).
11. Journal of Gene Medicine. Gene Therapy Clinical Trials Worldwide. Available online: http://www.abedia.com/wiley/index.html (accessed on 3 October 2013).
12. Zerboni, L.; Arvin, A. Investigation of varicella-zoster virus neurotropism and neurovirulence using scid mouse-human drg xenografts. *J. Neurovirol.* **2011**, *17*, 570–577.
13. Bertke, A.S.; Patel, A.; Krause, P.R. Herpes simplex virus latency-associated transcript sequence downstream of the promoter influences type-specific reactivation and viral neurotropism. *J. Virol.* **2007**, *81*, 6605–6613.
14. Lachmann, R. Herpes simplex virus-based vectors. *Int. J. Exp. Pathol.* **2004**, *85*, 177–190.

15. Muik, A.; Kneiske, I.; Werbizki, M.; Wilflingseder, D.; Giroglou, T.; Ebert, O.; Kraft, A.; Dietrich, U.; Zimmer, G.; Momma, S.; *et al.* Pseudotyping vesicular stomatitis virus with lymphocytic choriomeningitis virus glycoproteins enhances infectivity for glioma cells and minimizes neurotropism. *J. Virol.* **2011**, *85*, 5679–5684.
16. Kim, S.H.; Xiao, S.; Shive, H.; Collins, P.L.; Samal, S.K. Replication, neurotropism, and pathogenicity of avian paramyxovirus serotypes 1–9 in chickens and ducks. *PLoS One* **2012**, *7*, e34927.
17. Yumitori, K.; Handa, H.; Yamashita, J.; Suda, K.; Otsuka, S.; Shimizu, Y. Treatment of malignant glioma with mumps virus (author's transl). *No Shinkei Geka* **1982**, *10*, 143–147.
18. Short, M.P.; Choi, B.C.; Lee, J.K.; Malick, A.; Breakefield, X.O.; Martuza, R.L. Gene delivery to glioma cells in rat brain by grafting of a retrovirus packaging cell line. *J. Neurosci. Res.* **1990**, *27*, 427–439.
19. Culver, K.W.; Ram, Z.; Wallbridge, S.; Ishii, H.; Oldfield, E.H.; Blaese, R.M. *In vivo* gene transfer with retroviral vector-producer cells for treatment of experimental brain tumors. *Science* **1992**, *256*, 1550–1552.
20. Martuza, R.L.; Malick, A.; Markert, J.M.; Ruffner, K.L.; Coen, D.M. Experimental therapy of human glioma by means of a genetically engineered virus mutant. *Science* **1991**, *252*, 854–856.
21. Dachs, G.U.; Tupper, J.; Tozer, G.M. From bench to bedside for gene-directed enzyme prodrug therapy of cancer. *Anticancer Drugs* **2005**, *16*, 349–359.
22. Duarte, S.; Carle, G.; Faneca, H.; de Lima, M.C.; Pierrefite-Carle, V. Suicide gene therapy in cancer: Where do we stand now? *Cancer Lett.* **2012**, *324*, 160–170.
23. Kaliberov, S.A.; Market, J.M.; Gillespie, G.Y.; Krendelchtchikova, V.; Della Manna, D.; Sellers, J.C.; Kaliberova, L.N.; Black, M.E.; Buchsbaum, D.J. Mutation of *Escherichia coli* cytosine deaminase significantly enhances molecular chemotherapy of human glioma. *Gene Ther.* **2007**, *14*, 1111–1119.
24. Moolten, F.L. Tumor chemosensitivity conferred by inserted herpes thymidine kinase genes: Paradigm for a prospective cancer control strategy. *Cancer Res.* **1986**, *46*, 5276–5281.
25. Valerie, K.; Hawkins, W.; Farnsworth, J.; Schmidt-Ullrich, R.; Lin, P.S.; Amir, C.; Feden, J. Substantially improved *in vivo* radiosensitization of rat glioma with mutant hsv-tk and acyclovir. *Cancer Gene Ther.* **2001**, *8*, 3–8.
26. Ezzeddine, Z.D.; Martuza, R.L.; Platika, D.; Short, M.P.; Malick, A.; Choi, B.; Breakefield, X.O. Selective killing of glioma cells in culture and *in vivo* by retrovirus transfer of the herpes simplex virus thymidine kinase gene. *New Biol.* **1991**, *3*, 608–614.
27. Maeda, M.; Moriuchi, S.; Sano, A.; Yoshimine, T. New drug delivery system for water-soluble drugs using silicone and its usefulness for local treatment: Application of gcv-silicone to gcv/hsv-tk gene therapy for brain tumor. *J. Control. Release* **2002**, *84*, 15–25.
28. Valerie, K.; Brust, D.; Farnsworth, J.; Amir, C.; Taher, M.M.; Hershey, C.; Feden, J. Improved radiosensitization of rat glioma cells with adenovirus-expressed mutant herpes simplex virus-thymidine kinase in combination with acyclovir. *Cancer Gene Ther.* **2000**, *7*, 879–884.
29. Namba, H.; Tagawa, M.; Miyagawa, T.; Iwadate, Y.; Sakiyama, S. Treatment of rat experimental brain tumors by herpes simplex virus thymidine kinase gene-transduced allogeneic tumor cells and ganciclovir. *Cancer Gene Ther.* **2000**, *7*, 947–953.

30. Mori, K.; Iwata, J.; Miyazaki, M.; Osada, H.; Tange, Y.; Yamamoto, T.; Aiko, Y.; Tamura, M.; Shiroishi, T. Bystander killing effect of tymidine kinase gene-transduced adult bone marrow stromal cells with ganciclovir on malignant glioma cells. *Neurol. Med. Chir. (Tokyo)* **2010**, *50*, 545–553.
31. Freeman, S.M.; Abboud, C.N.; Whartenby, K.A.; Packman, C.H.; Koeplin, D.S.; Moolten, F.L.; Abraham, G.N. The “bystander effect”: Tumor regression when a fraction of the tumor mass is genetically modified. *Cancer Res.* **1993**, *53*, 5274–5283.
32. Bi, W.L.; Parysek, L.M.; Warnick, R.; Stambrook, P.J. *In vitro* evidence that metabolic cooperation is responsible for the bystander effect observed with hsv tk retroviral gene therapy. *Hum. Gene Ther.* **1993**, *4*, 725–731.
33. Fick, J.; Barker, F.G., 2nd; Dazin, P.; Westphale, E.M.; Beyer, E.C.; Israel, M.A. The extent of heterocellular communication mediated by gap junctions is predictive of bystander tumor cytotoxicity *in vitro*. *Proc. Natl. Acad. Sci. USA* **1995**, *92*, 11071–11075.
34. Kim, S.H.; Kim, J.H.; Kolozsvary, A.; Brown, S.L.; Freytag, S.O. Preferential radiosensitization of 9l glioma cells transduced with hsv-tk gene by acyclovir. *J. Neurooncol.* **1997**, *33*, 189–194.
35. Wahlfors, T.; Hakkarainen, T.; Janne, J.; Alhonen, L.; Wahlfors, J. *In vivo* enhancement of herpes simplex virus thymidine kinase/ganciclovir cancer gene therapy with polyamine biosynthesis inhibition. *Int. J. Cancer* **2006**, *118*, 2907–2910.
36. Rainov, N.G.; Fels, C.; Droege, J.W.; Schafer, C.; Kramm, C.M.; Chou, T.C. Temozolomide enhances herpes simplex virus thymidine kinase/ganciclovir therapy of malignant glioma. *Cancer Gene Ther.* **2001**, *8*, 662–668.
37. Rainov, N.G. A phase III clinical evaluation of herpes simplex virus type 1 thymidine kinase and ganciclovir gene therapy as an adjuvant to surgical resection and radiation in adults with previously untreated glioblastoma multiforme. *Hum. Gene Ther.* **2000**, *11*, 2389–2401.
38. Goodman, J.C.; Trask, T.W.; Chen, S.H.; Woo, S.L.; Grossman, R.G.; Carey, K.D.; Hubbard, G.B.; Carrier, D.A.; Rajagopalan, S.; Aguilar-Cordova, E.; *et al.* Adenoviral-mediated thymidine kinase gene transfer into the primate brain followed by systemic ganciclovir: Pathologic, radiologic, and molecular studies. *Hum. Gene Ther.* **1996**, *7*, 1241–1250.
39. Dewey, R.A.; Morrissey, G.; Cowsill, C.M.; Stone, D.; Bolognani, F.; Dodd, N.J.; Southgate, T.D.; Klatzmann, D.; Lassmann, H.; Castro, M.G.; *et al.* Chronic brain inflammation and persistent herpes simplex virus 1 thymidine kinase expression in survivors of syngeneic glioma treated by adenovirus-mediated gene therapy: Implications for clinical trials. *Nat. Med.* **1999**, *5*, 1256–1263.
40. Prados, M.D.; McDermott, M.; Chang, S.M.; Wilson, C.B.; Fick, J.; Culver, K.W.; van Gilder, J.; Keles, G.E.; Spence, A.; Berger, M. Treatment of progressive or recurrent glioblastoma multiforme in adults with herpes simplex virus thymidine kinase gene vector-producer cells followed by intravenous ganciclovir administration: A phase I/II multi-institutional trial. *J. Neurooncol.* **2003**, *65*, 269–278.
41. Germano, I.M.; Fable, J.; Gultekin, S.H.; Silvers, A. Adenovirus/herpes simplex-thymidine kinase/ganciclovir complex: Preliminary results of a phase I trial in patients with recurrent malignant gliomas. *J. Neurooncol.* **2003**, *65*, 279–289.

42. Assi, H.; Candolfi, M.; Baker, G.; Mineharu, Y.; Lowenstein, P.R.; Castro, M.G. Gene therapy for brain tumors: Basic developments and clinical implementation. *Neurosci. Lett.* **2012**, *527*, 71–77.
43. Adachi, Y.; Tamiya, T.; Ichikawa, T.; Terada, K.; Ono, Y.; Matsumoto, K.; Furuta, T.; Hamada, H.; Ohmoto, T. Experimental gene therapy for brain tumors using adenovirus-mediated transfer of cytosine deaminase gene and uracil phosphoribosyltransferase gene with 5-fluorocytosine. *Hum. Gene Ther.* **2000**, *11*, 77–89.
44. Bourbeau, D.; Lavoie, G.; Nalbantoglu, J.; Massie, B. Suicide gene therapy with an adenovirus expressing the fusion gene cd::Uprt in human glioblastomas: Different sensitivities correlate with p53 status. *J. Gene Med.* **2004**, *6*, 1320–1332.
45. Ostertag, D.; Amundson, K.K.; Lopez Espinoza, F.; Martin, B.; Buckley, T.; Galvao da Silva, A.P.; Lin, A.H.; Valenta, D.T.; Perez, O.D.; Ibanez, C.E.; *et al*. Brain tumor eradication and prolonged survival from intratumoral conversion of 5-fluorocytosine to 5-fluorouracil using a nonlytic retroviral replicating vector. *Neuro-oncology* **2012**, *14*, 145–159.
46. Ardiani, A.; Johnson, A.J.; Ruan, H.; Sanchez-Bonilla, M.; Serve, K.; Black, M.E. Enzymes to die for: Exploiting nucleotide metabolizing enzymes for cancer gene therapy. *Curr. Gene Ther.* **2012**, *12*, 77–91.
47. Tai, C.K.; Wang, W.; Lai, Y.H.; Logg, C.R.; Parker, W.B.; Li, Y.F.; Hong, J.S.; Sorscher, E.J.; Chen, T.C.; Kasahara, N. Enhanced efficiency of prodrug activation therapy by tumor-selective replicating retrovirus vectors armed with the escherichia coli purine nucleoside phosphorylase gene. *Cancer Gene Ther.* **2010**, *17*, 614–623.
48. Bharara, S.; Sorscher, E.J.; Gillespie, G.Y.; Lindsey, J.R.; Hong, J.S.; Curlee, K.V.; Allan, P.W.; Gadi, V.K.; Alexander, S.A.; Secrist, J.A., 3rd; *et al.* Antibiotic-mediated chemoprotection enhances adaptation of *E. coli* pnp for herpes simplex virus-based glioma therapy. *Hum. Gene Ther.* **2005**, *16*, 339–347.
49. Marais, R.; Spooner, R.A.; Stribbling, S.M.; Light, Y.; Martin, J.; Springer, C.J. A cell surface tethered enzyme improves efficiency in gene-directed enzyme prodrug therapy. *Nat. Biotechnol.* **1997**, *15*, 1373–1377.
50. Spooner, R.A.; Martin, J.; Friedlos, F.; Marais, R.; Springer, C.J. In suicide gene therapy, the site of subcellular localization of the activating enzyme is more important than the rate at which it activates prodrug. *Cancer Gene Ther.* **2000**, *7*, 1348–1356.
51. Cowen, R.L.; Williams, J.C.; Emery, S.; Blakey, D.; Darling, J.L.; Lowenstein, P.R.; Castro, M.G. Adenovirus vector-mediated delivery of the prodrug-converting enzyme carboxypeptidase g2 in a secreted or gpi-anchored form: High-level expression of this active conditional cytotoxic enzyme at the plasma membrane. *Cancer Gene Ther.* **2002**, *9*, 897–907.
52. Wei, M.X.; Tamiya, T.; Chase, M.; Boviatsis, E.J.; Chang, T.K.; Kowall, N.W.; Hochberg, F.H.; Waxman, D.J.; Breakefield, X.O.; Chiocca, E.A. Experimental tumor therapy in mice using the cyclophosphamide-activating cytochrome p450 2b1 gene. *Hum. Gene Ther.* **1994**, *5*, 969–978.
53. Chase, M.; Chung, R.Y.; Chiocca, E.A. An oncolytic viral mutant that delivers the cyp2b1 transgene and augments cyclophosphamide chemotherapy. *Nat. Biotechnol.* **1998**, *16*, 444–448.

54. Fulci, G.; Breymann, L.; Gianni, D.; Kurozomi, K.; Rhee, S.S.; Yu, J.; Kaur, B.; Louis, D.N.; Weissleder, R.; Caligiuri, M.A.; *et al.* Cyclophosphamide enhances glioma virotherapy by inhibiting innate immune responses. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 12873–12878.
55. Mut, M.; Sherman, J.H.; Shaffrey, M.E.; Schiff, D. Cintredekin besudotox in treatment of malignant glioma. *Expert Opin. Biol. Ther.* **2008**, *8*, 805–812.
56. Candolfi, M.; Xiong, W.; Yagiz, K.; Liu, C.; Muhammad, A.K.; Puntel, M.; Foulad, D.; Zadmehr, A.; Ahlzadeh, G.E.; Kroeger, K.M.; *et al.* Gene therapy-mediated delivery of targeted cytotoxins for glioma therapeutics. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 20021–20026.
57. Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **2008**, *455*, 1061–1068.
58. Ohgaki, H.; Kleihues, P. The definition of primary and secondary glioblastoma. *Clin. Cancer Res.* **2013**, *19*, 764–772.
59. Lang, F.F.; Yung, W.K.; Sawaya, R.; Tofilon, P.J. Adenovirus-mediated p53 gene therapy for human gliomas. *Neurosurgery* **1999**, *45*, 1093–1104.
60. Hong, Y.K.; Joe, Y.A.; Yang, Y.J.; Lee, K.S.; Son, B.C.; Jeun, S.S.; Chung, D.S.; Cho, K.K.; Park, C.K.; Kim, M.C.; *et al.* Potentials and limitations of adenovirus-p53 gene therapy for brain tumors. *J. Korean Med. Sci.* **2000**, *15*, 315–322.
61. Lang, F.F.; Yung, W.K.; Raju, U.; Libunao, F.; Terry, N.H.; Tofilon, P.J. Enhancement of radiosensitivity of wild-type p53 human glioma cells by adenovirus-mediated delivery of the p53 gene. *J. Neurosurg.* **1998**, *89*, 125–132.
62. Li, H.; Alonso-Vanegas, M.; Colicos, M.A.; Jung, S.S.; Lochmuller, H.; Sadikot, A.F.; Snipes, G.J.; Seth, P.; Karpati, G.; Nalbantoglu, J. Intracerebral adenovirus-mediated p53 tumor suppressor gene therapy for experimental human glioma. *Clin. Cancer Res.* **1999**, *5*, 637–642.
63. Biroccio, A.; Bufalo, D.D.; Ricca, A.; D'Angelo, C.; D'Orazi, G.; Sacchi, A.; Soddu, S.; Zupi, G. Increase of bcnu sensitivity by wt-p53 gene therapy in glioblastoma lines depends on the administration schedule. *Gene Ther.* **1999**, *6*, 1064–1072.
64. Kim, I.A.; Yang, Y.J.; Yoon, S.C.; Choi, I.B.; Kay, C.S.; Kwon, H.C.; Kim, C.M.; Joe, Y.A.; Kang, J.K.; Hong, Y.K. Potential of adenoviral p53 gene therapy and irradiation for the treatment of malignant gliomas. *Int. J. Oncol.* **2001**, *19*, 1041–1047.
65. Chen, B.; Timiryasova, T.M.; Andres, M.L.; Kajioka, E.H.; Dutta-Roy, R.; Gridley, D.S.; Fodor, I. Evaluation of combined vaccinia virus-mediated antitumor gene therapy with p53, il-2, and il-12 in a glioma model. *Cancer Gene Ther.* **2000**, *7*, 1437–1447.
66. Lang, F.F.; Bruner, J.M.; Fuller, G.N.; Aldape, K.; Prados, M.D.; Chang, S.; Berger, M.S.; McDermott, M.W.; Kunwar, S.M.; Junck, L.R.; *et al.* Phase i trial of adenovirus-mediated p53 gene therapy for recurrent glioma: Biological and clinical results. *J. Clin. Oncol.* **2003**, *21*, 2508–2518.
67. Chintala, S.K.; Fueyo, J.; Gomez-Manzano, C.; Venkaiah, B.; Bjerkvig, R.; Yung, W.K.; Sawaya, R.; Kyritsis, A.P.; Rao, J.S. Adenovirus-mediated p16/cdkn2 gene transfer suppresses glioma invasion *in vitro*. *Oncogene* **1997**, *15*, 2049–2057.
68. Park, K.H.; Lee, J.; Yoo, C.G.; Kim, Y.W.; Han, S.K.; Shim, Y.S.; Kim, S.K.; Wang, K.C.; Cho, B.K.; Lee, C.T. Application of p27 gene therapy for human malignant glioma potentiated by using mutant p27. *J. Neurosurg.* **2004**, *101*, 505–510.

69. Dunn, G.P.; Rinne, M.L.; Wykosky, J.; Genovese, G.; Quayle, S.N.; Dunn, I.F.; Agarwalla, P.K.; Chheda, M.G.; Campos, B.; Wang, A.; *et al.* Emerging insights into the molecular and cellular basis of glioblastoma. *Genes Dev.* **2012**, *26*, 756–784.
70. Davies, M.A.; Lu, Y.; Sano, T.; Fang, X.; Tang, P.; LaPushin, R.; Koul, D.; Bookstein, R.; Stokoe, D.; Yung, W.K.; *et al.* Adenoviral transgene expression of mmac/pten in human glioma cells inhibits akt activation and induces anoikis. *Cancer Res.* **1998**, *58*, 5285–5290.
71. Lu, W.; Zhou, X.; Hong, B.; Liu, J.; Yue, Z. Suppression of invasion in human u87 glioma cells by adenovirus-mediated co-transfer of timp-2 and pten gene. *Cancer Lett.* **2004**, *214*, 205–213.
72. Abe, T.; Terada, K.; Wakimoto, H.; Inoue, R.; Tyminski, E.; Bookstein, R.; Basilion, J.P.; Chiocca, E.A. Pten decreases *in vivo* vascularization of experimental gliomas in spite of proangiogenic stimuli. *Cancer Res.* **2003**, *63*, 2300–2305.
73. Vervoorts, J.; Luscher, B. Post-translational regulation of the tumor suppressor p27(kip1). *Cell Mol. Life Sci.* **2008**, *65*, 3255–3264.
74. Schiappacassi, M.; Lovat, F.; Canzonieri, V.; Belletti, B.; Berton, S.; di Stefano, D.; Vecchione, A.; Colombatti, A.; Baldassarre, G. P27kip1 expression inhibits glioblastoma growth, invasion, and tumor-induced neoangiogenesis. *Mol. Cancer Ther.* **2008**, *7*, 1164–1175.
75. Qin, X.Q.; Beckham, C.; Brown, J.L.; Lukashev, M.; Barsoum, J. Human and mouse ifn-beta gene therapy exhibits different anti-tumor mechanisms in mouse models. *Mol. Ther.* **2001**, *4*, 356–364.
76. Qin, X.Q.; Tao, N.; Dergay, A.; Moy, P.; Fawell, S.; Davis, A.; Wilson, J.M.; Barsoum, J. Interferon-beta gene therapy inhibits tumor formation and causes regression of established tumors in immune-deficient mice. *Proc. Natl. Acad. Sci. USA* **1998**, *95*, 14411–14416.
77. Chiocca, E.A.; Smith, K.M.; McKinney, B.; Palmer, C.A.; Rosenfeld, S.; Lillehei, K.; Hamilton, A.; DeMasters, B.K.; Judy, K.; Kirn, D. A phase I trial of Ad.Hifn-beta gene therapy for glioma. *Mol. Ther.* **2008**, *16*, 618–626.
78. Enderlin, M.; Kleinmann, E.V.; Struyf, S.; Buracchi, C.; Vecchi, A.; Kinscherf, R.; Kiessling, F.; Paschek, S.; Sozzani, S.; Rommelaere, J.; *et al.* TNF-alpha and the ifn-gamma-inducible protein 10 (ip-10/cxcl-10) delivered by parvoviral vectors act in synergy to induce antitumor effects in mouse glioblastoma. *Cancer Gene Ther.* **2009**, *16*, 149–160.
79. Hellums, E.K.; Markert, J.M.; Parker, J.N.; He, B.; Perbal, B.; Roizman, B.; Whitley, R.J.; Langford, C.P.; Bharara, S.; Gillespie, G.Y. Increased efficacy of an interleukin-12-secreting herpes simplex virus in a syngeneic intracranial murine glioma model. *Neuro-oncology* **2005**, *7*, 213–224.
80. Chiu, T.L.; Wang, M.J.; Su, C.C. The treatment of glioblastoma multiforme through activation of microglia and trail induced by raav2-mediated il-12 in a syngeneic rat model. *J. Biomed. Sci.* **2012**, doi:10.1186/1423-0127-19-45.
81. Markert, J.M.; Cody, J.J.; Parker, J.N.; Coleman, J.M.; Price, K.H.; Kern, E.R.; Quenelle, D.C.; Lakeman, A.D.; Schoeb, T.R.; Palmer, C.A.; *et al.* Preclinical evaluation of a genetically engineered herpes simplex virus expressing interleukin-12. *J. Virol.* **2012**, *86*, 5304–5313.
82. Tanaka, T.; Cao, Y.; Folkman, J.; Fine, H.A. Viral vector-targeted antiangiogenic gene therapy utilizing an angiostatin complementary DNA. *Cancer Res.* **1998**, *58*, 3362–3369.

83. Ma, H.I.; Lin, S.Z.; Chiang, Y.H.; Li, J.; Chen, S.L.; Tsao, Y.P.; Xiao, X. Intratumoral gene therapy of malignant brain tumor in a rat model with angiostatin delivered by adeno-associated viral (aav) vector. *Gene Ther.* **2002**, *9*, 2–11.
84. Szentirmai, O.; Baker, C.H.; Bullain, S.S.; Lin, N.; Takahashi, M.; Folkman, J.; Mulligan, R.C.; Carter, B.S. Successful inhibition of intracranial human glioblastoma multiforme xenograft growth via systemic adenoviral delivery of soluble endostatin and soluble vascular endothelial growth factor receptor-2: Laboratory investigation. *J. Neurosurg.* **2008**, *108*, 979–988.
85. Hardcastle, J.; Kurozumi, K.; Dmitrieva, N.; Sayers, M.P.; Ahmad, S.; Waterman, P.; Weissleder, R.; Chiocca, E.A.; Kaur, B. Enhanced antitumor efficacy of vasculostatin (vstat120) expressing oncolytic hsv-1. *Mol. Ther.* **2010**, *18*, 285–294.
86. Yoo, J.Y.; Haseley, A.; Bratasz, A.; Chiocca, E.A.; Zhang, J.; Powell, K.; Kaur, B. Antitumor efficacy of 34.5enve: A transcriptionally retargeted and “vstat120”-expressing oncolytic virus. *Mol. Ther.* **2012**, *20*, 287–297.
87. Haseley, A.; Alvarez-Breckenridge, C.; Chaudhury, A.R.; Kaur, B. Advances in oncolytic virus therapy for glioma. *Recent Pat. CNS Drug Discov.* **2009**, *4*, 1–13.
88. Kuriyama, N.; Kuriyama, H.; Julin, C.M.; Lamborn, K.; Israel, M.A. Pretreatment with protease is a useful experimental strategy for enhancing adenovirus-mediated cancer gene therapy. *Hum. Gene Ther.* **2000**, *11*, 2219–2230.
89. Dmitrieva, N.; Yu, L.; Viapiano, M.; Cripe, T.P.; Chiocca, E.A.; Glorioso, J.C.; Kaur, B. Chondroitinase abc i-mediated enhancement of oncolytic virus spread and antitumor efficacy. *Clin. Cancer Res.* **2011**, *17*, 1362–1372.
90. Chiocca, E.A. Oncolytic viruses. *Nat. Rev. Cancer* **2002**, *2*, 938–950.
91. Selznick, L.A.; Shamji, M.F.; Fecci, P.; Gromeier, M.; Friedman, A.H.; Sampson, J. Molecular strategies for the treatment of malignant glioma—Genes, viruses, and vaccines. *Neurosurg. Rev.* **2008**, *31*, 141–155.
92. Allen, C.; Vongpunsawad, S.; Nakamura, T.; James, C.D.; Schroeder, M.; Cattaneo, R.; Giannini, C.; Krempski, J.; Peng, K.W.; Goble, J.M.; *et al.* Retargeted oncolytic measles strains entering via the egfrviii receptor maintain significant antitumor activity against gliomas with increased tumor specificity. *Cancer Res.* **2006**, *66*, 11840–11850.
93. Grandi, P.; Fernandez, J.; Szentirmai, O.; Carter, R.; Gianni, D.; Sena-Esteves, M.; Breakefield, X.O. Targeting hsv-1 virions for specific binding to epidermal growth factor receptor-viii-bearing tumor cells. *Cancer Gene Ther.* **2010**, *17*, 655–663.
94. Uchida, H.; Marzulli, M.; Nakano, K.; Goins, W.F.; Chan, J.; Hong, C.S.; Mazzacurati, L.; Yoo, J.Y.; Haseley, A.; Nakashima, H.; *et al.* Effective treatment of an orthotopic xenograft model of human glioblastoma using an egfr-retargeted oncolytic herpes simplex virus. *Mol. Ther.* **2013**, *21*, 561–569.
95. Tobias, A.; Ahmed, A.; Moon, K.S.; Lesniak, M.S. The art of gene therapy for glioma: A review of the challenging road to the bedside. *J. Neurol. Neurosurg. Psychiatry* **2013**, *84*, 213–222.
96. Terada, K.; Wakimoto, H.; Tyminski, E.; Chiocca, E.A.; Saeki, Y. Development of a rapid method to generate multiple oncolytic hsv vectors and their *in vivo* evaluation using syngeneic mouse tumor models. *Gene Ther.* **2006**, *13*, 705–714.

97. Markert, J.M.; Medlock, M.D.; Rabkin, S.D.; Gillespie, G.Y.; Todo, T.; Hunter, W.D.; Palmer, C.A.; Feigenbaum, F.; Tornatore, C.; Tufaro, F.; *et al.* Conditionally replicating herpes simplex virus mutant, g207 for the treatment of malignant glioma: Results of a phase I trial. *Gene Ther.* **2000**, *7*, 867–874.
98. Markert, J.M.; Liechty, P.G.; Wang, W.; Gaston, S.; Braz, E.; Karrasch, M.; Nabors, L.B.; Markiewicz, M.; Lakeman, A.D.; Palmer, C.A.; *et al.* Phase ib trial of mutant herpes simplex virus g207 inoculated pre-and post-tumor resection for recurrent gbm. *Mol. Ther.* **2009**, *17*, 199–207.
99. Papanastassiou, V.; Rampling, R.; Fraser, M.; Petty, R.; Hadley, D.; Nicoll, J.; Harland, J.; Mabbs, R.; Brown, M. The potential for efficacy of the modified (icp 34.5(−)) herpes simplex virus hsv1716 following intratumoural injection into human malignant glioma: A proof of principle study. *Gene Ther.* **2002**, *9*, 398–406.
100. Harrow, S.; Papanastassiou, V.; Harland, J.; Mabbs, R.; Petty, R.; Fraser, M.; Hadley, D.; Patterson, J.; Brown, S.M.; Rampling, R. Hsv1716 injection into the brain adjacent to tumour following surgical resection of high-grade glioma: Safety data and long-term survival. *Gene Ther.* **2004**, *11*, 1648–1658.
101. Kanai, R.; Zaupa, C.; Sgubin, D.; Antoszczyk, S.J.; Martuza, R.L.; Wakimoto, H.; Rabkin, S.D. Effect of gamma34.5 deletions on oncolytic herpes simplex virus activity in brain tumors. *J. Virol.* **2012**, *86*, 4420–4431.
102. Kambara, H.; Okano, H.; Chiocca, E.A.; Saeki, Y. An oncolytic hsv-1 mutant expressing icp34.5 under control of a nestin promoter increases survival of animals even when symptomatic from a brain tumor. *Cancer Res.* **2005**, *65*, 2832–2839.
103. Hardcastle, J.; Kurozumi, K.; Chiocca, E.A.; Kaur, B. Oncolytic viruses driven by tumor-specific promoters. *Curr. Cancer Drug Targets* **2007**, *7*, 181–189.
104. Nandi, S.; Lesniak, M.S. Adenoviral virotherapy for malignant brain tumors. *Expert Opin. Biol. Ther.* **2009**, *9*, 737–747.
105. Bischoff, J.R.; Kirn, D.H.; Williams, A.; Heise, C.; Horn, S.; Muna, M.; Ng, L.; Nye, J.A.; Sampson-Johannes, A.; Fattaey, A.; *et al.* An adenovirus mutant that replicates selectively in p53-deficient human tumor cells. *Science* **1996**, *274*, 373–376.
106. Geoerger, B.; Grill, J.; Opolon, P.; Morizet, J.; Aubert, G.; Terrier-Lacombe, M.J.; Bressac De-Paillerets, B.; Barrois, M.; Feunteun, J.; Kirn, D.H.; *et al.* Oncolytic activity of the e1b-55 kda-deleted adenovirus onyx-015 is independent of cellular p53 status in human malignant glioma xenografts. *Cancer Res.* **2002**, *62*, 764–772.
107. Chiocca, E.A.; Abbed, K.M.; Tatter, S.; Louis, D.N.; Hochberg, F.H.; Barker, F.; Kracher, J.; Grossman, S.A.; Fisher, J.D.; Carson, K.; *et al.* A phase i open-label, dose-escalation, multi-institutional trial of injection with an e1b-attenuated adenovirus, onyx-015, into the peritumoral region of recurrent malignant gliomas, in the adjuvant setting. *Mol. Ther.* **2004**, *10*, 958–966.
108. Fueyo, J.; Gomez-Manzano, C.; Alemany, R.; Lee, P.S.; McDonnell, T.J.; Mitlianga, P.; Shi, Y.X.; Levin, V.A.; Yung, W.K.; Kyritsis, A.P. A mutant oncolytic adenovirus targeting the rb pathway produces anti-glioma effect *in vivo*. *Oncogene* **2000**, *19*, 2–12.

109. Suzuki, K.; Fueyo, J.; Krasnykh, V.; Reynolds, P.N.; Curiel, D.T.; Alemany, R. A conditionally replicative adenovirus with enhanced infectivity shows improved oncolytic potency. *Clin. Cancer Res.* **2001**, *7*, 120–126.
110. Fueyo, J.; Alemany, R.; Gomez-Manzano, C.; Fuller, G.N.; Khan, A.; Conrad, C.A.; Liu, T.J.; Jiang, H.; Lemoine, M.G.; Suzuki, K.; *et al.* Preclinical characterization of the antiglioma activity of a tropism-enhanced adenovirus targeted to the retinoblastoma pathway. *J. Natl. Cancer Inst.* **2003**, *95*, 652–660.
111. Auffinger, B.; Ahmed, A.U.; Lesniak, M.S. Oncolytic virotherapy for malignant glioma: Translating laboratory insights into clinical practice. *Front. Oncol.* **2013**, doi:10.3389/fonc.2013.00032.
112. Ulasov, I.V.; Zhu, Z.B.; Tyler, M.A.; Han, Y.; Rivera, A.A.; Khramtsov, A.; Curiel, D.T.; Lesniak, M.S. Survivin-driven and fiber-modified oncolytic adenovirus exhibits potent antitumor activity in established intracranial glioma. *Hum. Gene Ther.* **2007**, *18*, 589–602.
113. Nandi, S.; Ulasov, I.V.; Tyler, M.A.; Sugihara, A.Q.; Molinero, L.; Han, Y.; Zhu, Z.B.; Lesniak, M.S. Low-dose radiation enhances survivin-mediated virotherapy against malignant glioma stem cells. *Cancer Res.* **2008**, *68*, 5778–5784.
114. Zemp, F.J.; Corredor, J.C.; Lun, X.; Muruve, D.A.; Forsyth, P.A. Oncolytic viruses as experimental treatments for malignant gliomas: Using a scourge to treat a devil. *Cytokine Growth Factor Rev.* **2010**, *21*, 103–117.
115. Ochiai, H.; Campbell, S.A.; Archer, G.E.; Chewning, T.A.; Dragunsky, E.; Ivanov, A.; Gromeier, M.; Sampson, J.H. Targeted therapy for glioblastoma multiforme neoplastic meningitis with intrathecal delivery of an oncolytic recombinant poliovirus. *Clin. Cancer Res.* **2006**, *12*, 1349–1354.
116. Pasquinucci, G. Possible effect of measles on leukaemia. *Lancet* **1971**, *1*, 136.
117. Bluming, A.Z.; Ziegler, J.L. Regression of burkitt's lymphoma in association with measles infection. *Lancet* **1971**, *2*, 105–106.
118. Allen, C.; Paraskevakou, G.; Iankov, I.; Giannini, C.; Schroeder, M.; Sarkaria, J.; Schroeder, M.; Puri, R.K.; Russell, S.J.; Galanis, E. Interleukin-13 displaying retargeted oncolytic measles virus strains have significant activity against gliomas with improved specificity. *Mol. Ther.* **2008**, *16*, 1556–1564.
119. Paraskevakou, G.; Allen, C.; Nakamura, T.; Zollman, P.; James, C.D.; Peng, K.W.; Schroeder, M.; Russell, S.J.; Galanis, E. Epidermal growth factor receptor (egfr)-retargeted measles virus strains effectively target egfr- or egfrviii expressing gliomas. *Mol. Ther.* **2007**, *15*, 677–686.
120. Phuong, L.K.; Allen, C.; Peng, K.W.; Giannini, C.; Greiner, S.; TenEyck, C.J.; Mishra, P.K.; Macura, S.I.; Russell, S.J.; Galanis, E.C. Use of a vaccine strain of measles virus genetically engineered to produce carcinoembryonic antigen as a novel therapeutic agent against glioblastoma multiforme. *Cancer Res.* **2003**, *63*, 2462–2469.
121. Myers, R.; Harvey, M.; Kaufmann, T.J.; Greiner, S.M.; Krempski, J.W.; Raffel, C.; Shelton, S.E.; Soeffker, D.; Zollman, P.; Federspiel, M.J.; *et al.* Toxicology study of repeat intracerebral administration of a measles virus derivative producing carcinoembryonic antigen in rhesus macaques in support of a phase I/II clinical trial for patients with recurrent gliomas. *Hum. Gene Ther.* **2008**, *19*, 690–698.

122. Allen, C.; Opyrchal, M.; Aderca, I.; Schroeder, M.A.; Sarkaria, J.N.; Domingo, E.; Federspiel, M.J.; Galanis, E. Oncolytic measles virus strains have significant antitumor activity against glioma stem cells. *Gene Ther.* **2013**, *20*, 444–449.
123. Chiocca, E.A.; Aghi, M.; Fulci, G. Viral therapy for glioblastoma. *Cancer J.* **2003**, *9*, 167–179.
124. Allen, C.; Paraskevakou, G.; Liu, C.; Iankov, I.D.; Msaouel, P.; Zollman, P.; Myers, R.; Peng, K.W.; Russell, S.J.; Galanis, E. Oncolytic measles virus strains in the treatment of gliomas. *Expert Opin. Biol. Ther.* **2008**, *8*, 213–220.
125. Kurozumi, K.; Hardcastle, J.; Thakur, R.; Yang, M.; Christoforidis, G.; Fulci, G.; Hochberg, F.H.; Weissleder, R.; Carson, W.; Chiocca, E.A.; *et al.* Effect of tumor microenvironment modulation on the efficacy of oncolytic virus therapy. *J. Natl. Cancer Inst.* **2007**, *99*, 1768–1781.
126. Sena-Esteves, M.; Hampl, J.A.; Camp, S.M.; Breakefield, X.O. Generation of stable retrovirus packaging cell lines after transduction with herpes simplex virus hybrid amplicon vectors. *J. Gene Med.* **2002**, *4*, 229–239.
127. Bjerkvig, R.; Read, T.A.; Vajkoczy, P.; Aebischer, P.; Pralong, W.; Platt, S.; Melvik, J.E.; Hagen, A.; Dornish, M. Cell therapy using encapsulated cells producing endostatin. *Acta Neurochir. Suppl.* **2003**, *88*, 137–141.
128. Aboody, K.S.; Brown, A.; Rainov, N.G.; Bower, K.A.; Liu, S.; Yang, W.; Small, J.E.; Herrlinger, U.; Ourednik, V.; Black, P.M.; *et al.* Neural stem cells display extensive tropism for pathology in adult brain: Evidence from intracranial gliomas. *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 12846–12851.
129. Aboody, K.S.; Najbauer, J.; Danks, M.K. Stem and progenitor cell-mediated tumor selective gene therapy. *Gene Ther.* **2008**, *15*, 739–752.
130. Gage, F.H. Mammalian neural stem cells. *Science* **2000**, *287*, 1433–1438.
131. Binello, E.; Germano, I.M. Stem cells as therapeutic vehicles for the treatment of high-grade gliomas. *Neuro-oncology* **2012**, *14*, 256–265.
132. Benedetti, S.; Pirola, B.; Pollo, B.; Magrassi, L.; Bruzzone, M.G.; Rigamonti, D.; Galli, R.; Selleri, S.; di Meco, F.; de Fraja, C.; *et al.* Gene therapy of experimental brain tumors using neural progenitor cells. *Nat. Med.* **2000**, *6*, 447–450.
133. Ahmed, A.U.; Alexiades, N.G.; Lesniak, M.S. The use of neural stem cells in cancer gene therapy: Predicting the path to the clinic. *Curr. Opin. Mol. Ther.* **2010**, *12*, 546–552.
134. Herrlinger, U.; Woiciechowski, C.; Sena-Esteves, M.; Aboody, K.S.; Jacobs, A.H.; Rainov, N.G.; Snyder, E.Y.; Breakefield, X.O. Neural precursor cells for delivery of replication-conditional hsv-1 vectors to intracerebral gliomas. *Mol. Ther.* **2000**, *1*, 347–357.
135. Tyler, M.A.; Ulasov, I.V.; Sonabend, A.M.; Nandi, S.; Han, Y.; Marler, S.; Roth, J.; Lesniak, M.S. Neural stem cells target intracranial glioma to deliver an oncolytic adenovirus *in vivo*. *Gene Ther.* **2009**, *16*, 262–278.
136. Ahmed, A.U.; Thaci, B.; Alexiades, N.G.; Han, Y.; Qian, S.; Liu, F.; Balyasnikova, I.V.; Ulasov, I.Y.; Aboody, K.S.; Lesniak, M.S. Neural stem cell-based cell carriers enhance therapeutic efficacy of an oncolytic adenovirus in an orthotopic mouse model of human glioblastoma. *Mol. Ther.* **2011**, *19*, 1714–1726.

137. Barresi, V.; Belluardo, N.; Sipione, S.; Mudo, G.; Cattaneo, E.; Condorelli, D.F. Transplantation of prodrug-converting neural progenitor cells for brain tumor therapy. *Cancer Gene Ther.* **2003**, *10*, 396–402.
138. Rath, P.; Shi, H.; Maruniak, J.A.; Litofsky, N.S.; Maria, B.L.; Kirk, M.D. Stem cells as vectors to deliver hsv/tk gene therapy for malignant gliomas. *Curr. Stem. Cell Res. Ther.* **2009**, *4*, 44–49.
139. Ito, S.; Natsume, A.; Shimato, S.; Ohno, M.; Kato, T.; Chansakul, P.; Wakabayashi, T.; Kim, S.U. Human neural stem cells transduced with ifn-beta and cytosine deaminase genes intensify bystander effect in experimental glioma. *Cancer Gene Ther.* **2010**, *17*, 299–306.
140. Ehtesham, M.; Kabos, P.; Kabosova, A.; Neuman, T.; Black, K.L.; Yu, J.S. The use of interleukin 12-secreting neural stem cells for the treatment of intracranial glioma. *Cancer Res.* **2002**, *62*, 5657–5663.
141. Yuan, X.; Hu, J.; Belladonna, M.L.; Black, K.L.; Yu, J.S. Interleukin-23-expressing bone marrow-derived neural stem-like cells exhibit antitumor activity against intracranial glioma. *Cancer Res.* **2006**, *66*, 2630–2638.
142. Shah, K.; Tung, C.H.; Yang, K.; Weissleder, R.; Breakefield, X.O. Inducible release of trail fusion proteins from a proapoptotic form for tumor therapy. *Cancer Res.* **2004**, *64*, 3236–3242.
143. Shah, K.; Bureau, E.; Kim, D.E.; Yang, K.; Tang, Y.; Weissleder, R.; Breakefield, X.O. Glioma therapy and real-time imaging of neural precursor cell migration and tumor regression. *Ann. Neurol.* **2005**, *57*, 34–41.
144. Hingtgen, S.; Ren, X.; Terwilliger, E.; Classon, M.; Weissleder, R.; Shah, K. Targeting multiple pathways in gliomas with stem cell and viral delivered s-trail and temozolomide. *Mol. Cancer Ther.* **2008**, *7*, 3575–3585.
145. Bagci-Onder, T.; Wakimoto, H.; Anderegg, M.; Cameron, C.; Shah, K. A dual pi3k/mtor inhibitor, pi-103, cooperates with stem cell-delivered trail in experimental glioma models. *Cancer Res.* **2011**, *71*, 154–163.
146. Balyasnikova, I.V.; Ferguson, S.D.; Han, Y.; Liu, F.; Lesniak, M.S. Therapeutic effect of neural stem cells expressing trail and bortezomib in mice with glioma xenografts. *Cancer Lett.* **2011**, *310*, 148–159.
147. Lorico, A.; Mercapide, J.; Solodushko, V.; Alexeyev, M.; Fodstad, O.; Rappa, G. Primary neural stem/progenitor cells expressing endostatin or cytochrome p450 for gene therapy of glioblastoma. *Cancer Gene Ther.* **2008**, *15*, 605–615.
148. Van Eekelen, M.; Sasportas, L.S.; Kasmieh, R.; Yip, S.; Figueiredo, J.L.; Louis, D.N.; Weissleder, R.; Shah, K. Human stem cells expressing novel tsp-1 variant have anti-angiogenic effect on brain tumors. *Oncogene* **2010**, *29*, 3185–3195.
149. Kim, S.K.; Cargioli, T.G.; Machluf, M.; Yang, W.; Sun, Y.; Al-Hashem, R.; Kim, S.U.; Black, P.M.; Carroll, R.S. Pex-producing human neural stem cells inhibit tumor growth in a mouse glioma model. *Clin. Cancer Res.* **2005**, *11*, 5965–5970.
150. Oreffo, R.O.; Cooper, C.; Mason, C.; Clements, M. Mesenchymal stem cells: Lineage, plasticity, and skeletal therapeutic potential. *Stem Cell Rev.* **2005**, *1*, 169–178.
151. Dvorak, H.F. Tumors: Wounds that do not heal. Similarities between tumor stroma generation and wound healing. *N. Engl. J. Med.* **1986**, *315*, 1650–1659.

152. Bexell, D.; Scheding, S.; Bengzon, J. Toward brain tumor gene therapy using multipotent mesenchymal stromal cell vectors. *Mol. Ther.* **2010**, *18*, 1067–1075.
153. Nakamura, K.; Ito, Y.; Kawano, Y.; Kurozumi, K.; Kobune, M.; Tsuda, H.; Bizen, A.; Honmou, O.; Niitsu, Y.; Hamada, H. Antitumor effect of genetically engineered mesenchymal stem cells in a rat glioma model. *Gene Ther.* **2004**, *11*, 1155–1164.
154. Chang, D.Y.; Yoo, S.W.; Hong, Y.; Kim, S.; Kim, S.J.; Yoon, S.H.; Cho, K.G.; Paek, S.H.; Lee, Y.D.; Kim, S.S.; *et al.* The growth of brain tumors can be suppressed by multiple transplantation of mesenchymal stem cells expressing cytosine deaminase. *Int. J. Cancer* **2010**, *127*, 1975–1983.
155. Choi, S.A.; Lee, J.Y.; Wang, K.C.; Phi, J.H.; Song, S.H.; Song, J.; Kim, S.K. Human adipose tissue-derived mesenchymal stem cells: Characteristics and therapeutic potential as cellular vehicles for prodrug gene therapy against brainstem gliomas. *Eur. J. Cancer* **2012**, *48*, 129–137.
156. Gunnarsson, S.; Bexell, D.; Svensson, A.; Siesjo, P.; Darabi, A.; Bengzon, J. Intratumoral il-7 delivery by mesenchymal stromal cells potentiates ifngamma-transduced tumor cell immunotherapy of experimental glioma. *J. Neuroimmunol.* **2010**, *218*, 140–144.
157. Ryu, C.H.; Park, S.H.; Park, S.A.; Kim, S.M.; Lim, J.Y.; Jeong, C.H.; Yoon, W.S.; Oh, W.I.; Sung, Y.C.; Jeun, S.S. Gene therapy of intracranial glioma using interleukin 12-secreting human umbilical cord blood-derived mesenchymal stem cells. *Hum. Gene Ther.* **2011**, *22*, 733–743.
158. Xu, G.; Jiang, X.D.; Xu, Y.; Zhang, J.; Huang, F.H.; Chen, Z.Z.; Zhou, D.X.; Shang, J.H.; Zou, Y.X.; Cai, Y.Q.; *et al.* Adenoviral-mediated interleukin-18 expression in mesenchymal stem cells effectively suppresses the growth of glioma in rats. *Cell Biol. Int.* **2009**, *33*, 466–474.
159. Kim, S.M.; Lim, J.Y.; Park, S.I.; Jeong, C.H.; Oh, J.H.; Jeong, M.; Oh, W.; Park, S.H.; Sung, Y.C.; Jeun, S.S. Gene therapy using trail-secreting human umbilical cord blood-derived mesenchymal stem cells against intracranial glioma. *Cancer Res.* **2008**, *68*, 9614–9623.
160. Menon, L.G.; Kelly, K.; Yang, H.W.; Kim, S.K.; Black, P.M.; Carroll, R.S. Human bone marrow-derived mesenchymal stromal cells expressing s-trail as a cellular delivery vehicle for human glioma therapy. *Stem Cells* **2009**, *27*, 2320–2330.
161. Kim, S.M.; Oh, J.H.; Park, S.A.; Ryu, C.H.; Lim, J.Y.; Kim, D.S.; Chang, J.W.; Oh, W.; Jeun, S.S. Irradiation enhances the tumor tropism and therapeutic potential of tumor necrosis factor-related apoptosis-inducing ligand-secreting human umbilical cord blood-derived mesenchymal stem cells in glioma therapy. *Stem Cells* **2010**, *28*, 2217–2228.
162. Kim, S.M.; Woo, J.S.; Jeong, C.H.; Ryu, C.H.; Lim, J.Y.; Jeun, S.S. Effective combination therapy for malignant glioma with trail-secreting mesenchymal stem cells and lipoxygenase inhibitor mk886. *Cancer Res.* **2012**, *72*, 4807–4817.
163. Ahmed, A.U.; Rolle, C.E.; Tyler, M.A.; Han, Y.; Sengupta, S.; Wainwright, D.A.; Balyasnikova, I.V.; Ulasov, I.V.; Lesniak, M.S. Bone marrow mesenchymal stem cells loaded with an oncolytic adenovirus suppress the anti-adenoviral immune response in the cotton rat model. *Mol. Ther.* **2010**, *18*, 1846–1856.
164. Ahmed, A.U.; Tyler, M.A.; Thaci, B.; Alexiades, N.G.; Han, Y.; Ulasov, I.V.; Lesniak, M.S. A comparative study of neural and mesenchymal stem cell-based carriers for oncolytic adenovirus in a model of malignant glioma. *Mol. Pharm.* **2011**, *8*, 1559–1572.

165. Yong, R.L.; Shinojima, N.; Fueyo, J.; Gumin, J.; Vecil, G.G.; Marini, F.C.; Bogler, O.; Andreeff, M.; Lang, F.F. Human bone marrow-derived mesenchymal stem cells for intravascular delivery of oncolytic adenovirus delta24-rgd to human gliomas. *Cancer Res.* **2009**, *69*, 8932–8940.
166. Keller, G. Embryonic stem cell differentiation: Emergence of a new era in biology and medicine. *Genes Dev.* **2005**, *19*, 1129–1155.
167. Germano, I.M.; Uzzaman, M.; Benveniste, R.J.; Zaurova, M.; Keller, G. Apoptosis in human glioblastoma cells produced using embryonic stem cell-derived astrocytes expressing tumor necrosis factor-related apoptosis-inducing ligand. *J. Neurosurg.* **2006**, *105*, 88–95.
168. Uzzaman, M.; Keller, G.; Germano, I.M. *In vivo* gene delivery by embryonic-stem-cell-derived astrocytes for malignant gliomas. *Neuro-oncology* **2009**, *11*, 102–108.
169. Germano, I.M.; Emdad, L.; Qadeer, Z.A.; Binello, E.; Uzzaman, M. Embryonic stem cell (esc)-mediated transgene delivery induces growth suppression, apoptosis and radiosensitization, and overcomes temozolomide resistance in malignant gliomas. *Cancer Gene Ther.* **2010**, *17*, 664–674.
170. Bak, X.Y.; Lam, D.H.; Yang, J.; Ye, K.; Wei, E.L.; Lim, S.K.; Wang, S. Human embryonic stem cell-derived mesenchymal stem cells as cellular delivery vehicles for prodrug gene therapy of glioblastoma. *Hum. Gene Ther.* **2011**, *22*, 1365–1377.
171. Zhao, Y.; Lam, D.H.; Yang, J.; Lin, J.; Tham, C.K.; Ng, W.H.; Wang, S. Targeted suicide gene therapy for glioma using human embryonic stem cell-derived neural stem cells genetically modified by baculoviral vectors. *Gene Ther.* **2012**, *19*, 189–200.
172. Nakashima, H.; Kaur, B.; Chiocca, E.A. Directing systemic oncolytic viral delivery to tumors via carrier cells. *Cytokine Growth Factor Rev.* **2010**, *21*, 119–126.
173. Tang, Y.; Shah, K.; Messerli, S.M.; Snyder, E.; Breakefield, X.; Weissleder, R. *In vivo* tracking of neural progenitor cell migration to glioblastomas. *Hum. Gene Ther.* **2003**, *14*, 1247–1254.
174. Reitz, M.; Demestre, M.; Sedlacik, J.; Meissner, H.; Fiehler, J.; Kim, S.U.; Westphal, M.; Schmidt, N.O. Intranasal delivery of neural stem/progenitor cells: A noninvasive passage to target intracerebral glioma. *Stem Cells Transl. Med.* **2012**, *1*, 866–873.
175. Assanah, M.; Lochhead, R.; Ogden, A.; Bruce, J.; Goldman, J.; Canoll, P. Glial progenitors in adult white matter are driven to form malignant gliomas by platelet-derived growth factor-expressing retroviruses. *J. Neurosci.* **2006**, *26*, 6781–6790.
176. Yoshida, J.; Mizuno, M. Clinical gene therapy for brain tumors. Liposomal delivery of anticancer molecule to glioma. *J. Neurooncol.* **2003**, *65*, 261–267.
177. Niidome, T.; Huang, L. Gene therapy progress and prospects: Nonviral vectors. *Gene Ther.* **2002**, *9*, 1647–1652.
178. Barbu, E.; Molnar, E.; Tsibouklis, J.; Gorecki, D.C. The potential for nanoparticle-based drug delivery to the brain: Overcoming the blood-brain barrier. *Expert Opin. Drug Deliv.* **2009**, *6*, 553–565.
179. Reszka, R.C.; Jacobs, A.; Voges, J. Liposome-mediated suicide gene therapy in humans. *Methods Enzymol.* **2005**, *391*, 200–208.

180. Yoshida, J.; Mizuno, M.; Fujii, M.; Kajita, Y.; Nakahara, N.; Hatano, M.; Saito, R.; Nobayashi, M.; Wakabayashi, T. Human gene therapy for malignant gliomas (glioblastoma multiforme and anaplastic astrocytoma) by *in vivo* transduction with human interferon beta gene using cationic liposomes. *Hum. Gene Ther.* **2004**, *15*, 77–86.
181. Sun, X.; Pang, Z.; Ye, H.; Qiu, B.; Guo, L.; Li, J.; Ren, J.; Qian, Y.; Zhang, Q.; Chen, J.; *et al.* Co-delivery of pegfp-htrail and paclitaxel to brain glioma mediated by an angiopep-conjugated liposome. *Biomaterials* **2012**, *33*, 916–924.
182. Ito, A.; Shinkai, M.; Bouhon, I.A.; Honda, H.; Kobayashi, T. Bystander-killing effect and cyclic induction of tnf-alpha gene under heat-inducible promoter gadd 153. *J. Biosci. Bioeng.* **2000**, *90*, 437–441.
183. Ito, A.; Shinkai, M.; Honda, H.; Kobayashi, T. Heat-inducible tnf-alpha gene therapy combined with hyperthermia using magnetic nanoparticles as a novel tumor-targeted therapy. *Cancer Gene Ther.* **2001**, *8*, 649–654.
184. Yin, T.; Wang, P.; Li, J.; Zheng, R.; Zheng, B.; Cheng, D.; Li, R.; Lai, J.; Shuai, X. Ultrasound-sensitive sirna-loaded nanobubbles formed by hetero-assembly of polymeric micelles and liposomes and their therapeutic effect in gliomas. *Biomaterials* **2013**, *34*, 4532–4543.
185. Godbey, W.T.; Wu, K.K.; Mikos, A.G. Poly(ethylenimine) and its role in gene delivery. *J. Control. Release* **1999**, *60*, 149–160.
186. Lu, X.; Ping, Y.; Xu, F.J.; Li, Z.H.; Wang, Q.Q.; Chen, J.H.; Yang, W.T.; Tang, G.P. Bifunctional conjugates comprising beta-cyclodextrin, polyethylenimine, and 5-fluoro-2'-deoxyuridine for drug delivery and gene transfer. *Bioconjug. Chem.* **2010**, *21*, 1855–1863.
187. Li, J.; Gu, B.; Meng, Q.; Yan, Z.; Gao, H.; Chen, X.; Yang, X.; Lu, W. The use of myristic acid as a ligand of polyethylenimine/DNA nanoparticles for targeted gene therapy of glioblastoma. *Nanotechnology* **2011**, *22*, 435101.
188. Zhan, C.; Meng, Q.; Li, Q.; Feng, L.; Zhu, J.; Lu, W. Cyclic rgd-polyethylene glycol-polyethylenimine for intracranial glioblastoma-targeted gene delivery. *Chem. Asian J.* **2012**, *7*, 91–96.
189. Dufes, C.; Uchegbu, I.F.; Schatzlein, A.G. Dendrimers in gene delivery. *Adv. Drug Deliv. Rev.* **2005**, *57*, 2177–2202.
190. Pettit, M.W.; Griffiths, P.; Ferruti, P.; Richardson, S.C. Poly(amidoamine) polymers: Soluble linear amphiphilic drug-delivery systems for genes, proteins and oligonucleotides. *Ther. Deliv.* **2011**, *2*, 907–917.
191. Han, L.; Zhang, A.; Wang, H.; Pu, P.; Jiang, X.; Kang, C.; Chang, J. Tat-bmps-pamam conjugates enhance therapeutic effect of small interference rna on u251 glioma cells *in vitro* and *in vivo*. *Hum. Gene Ther.* **2010**, *21*, 417–426.
192. Bai, C.Z.; Choi, S.; Nam, K.; An, S.; Park, J.S. Arginine modified pamam dendrimer for interferon beta gene delivery to malignant glioma. *Int. J. Pharm.* **2013**, *445*, 79–87.
193. Wang, H.; Su, W.; Wang, S.; Wang, X.; Liao, Z.; Kang, C.; Han, L.; Chang, J.; Wang, G.; Pu, P. Smart multifunctional core-shell nanospheres with drug and gene co-loaded for enhancing the therapeutic effect in a rat intracranial tumor model. *Nanoscale* **2012**, *4*, 6501–6508.
194. Kami, D.; Takeda, S.; Itakura, Y.; Gojo, S.; Watanabe, M.; Toyoda, M. Application of magnetic nanoparticles to gene delivery. *Int. J. Mol. Sci.* **2011**, *12*, 3705–3722.

195. Silva, A.C.; Oliveira, T.R.; Mamani, J.B.; Malheiros, S.M.; Malavolta, L.; Pavon, L.F.; Sibov, T.T.; Amaro, E., Jr.; Tannus, A.; Vidoto, E.L.; *et al.* Application of hyperthermia induced by superparamagnetic iron oxide nanoparticles in glioma treatment. *Int. J. Nanomed.* **2011**, *6*, 591–603.
196. Wankhede, M.; Bouras, A.; Kaluzova, M.; Hadjipanayis, C.G. Magnetic nanoparticles: An emerging technology for malignant brain tumor imaging and therapy. *Expert Rev. Clin. Pharmacol.* **2012**, *5*, 173–186.
197. Yun, J.; Sonabend, A.M.; Ulasov, I.V.; Kim, D.H.; Rozhkova, E.A.; Novosad, V.; Dashnaw, S.; Brown, T.; Canoll, P.; Bruce, J.N.; *et al.* A novel adenoviral vector labeled with superparamagnetic iron oxide nanoparticles for real-time tracking of viral delivery. *J. Clin. Neurosci.* **2012**, *19*, 875–880.
198. Kievit, F.M.; Veiseh, O.; Fang, C.; Bhattarai, N.; Lee, D.; Ellenbogen, R.G.; Zhang, M. Chlorotoxin labeled magnetic nanovectors for targeted gene delivery to glioma. *ACS Nano* **2010**, *4*, 4587–4594.
199. Veiseh, O.; Kievit, F.M.; Fang, C.; Mu, N.; Jana, S.; Leung, M.C.; Mok, H.; Ellenbogen, R.G.; Park, J.O.; Zhang, M. Chlorotoxin bound magnetic nanovector tailored for cancer cell targeting, imaging, and sirna delivery. *Biomaterials* **2010**, *31*, 8032–8042.

© 2013 by the authors; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/3.0/).

Uniwersytet Przyrodniczy we Wrocławiu

Wrocław University of Environmental and Life Sciences

https://bazawiedzy.upwr.edu.pl

Publikacja / Publication	Genetics, Genomics and Emerging Molecular Therapies of Pancreatic Cancer
Autorzy / authors	Liu Jakub, Mroczek Magdalena, Mach Anna, Stępień Maria, Aplas Angelika, Pronobis-Szczylik Bartosz, Bukowski Szymon, Mielczarek Magda, Gajewska Ewelina, Topolski Piotr
DOI wersji wydawcy / Published version DOI	http://dx.doi.org/10.3390/cancers15030779
Adres publikacji w Repozytorium URL / Publication address in Repository	https://bazawiedzy.upwr.edu.pl/info/article/UPWR5d83dfa585b14aed8b4a44e1e48c7119/
Data opublikowania w Repozytorium / Deposited in Repository on	2023-02-02
Rodzaj licencji / Type of licence	Uznanie Autorstwa / Attribution
Cytuj tę wersję / Cite this version	Liu J., Mroczek M., Mach A., Stępień M., Aplas A., Pronobis-Szczylik B., Bukowski S., et al., Genetics, Genomics and Emerging Molecular Therapies of Pancreatic Cancer, „Cancers”, 2023, t.15.

cancers

MDPI

Review

Genetics, Genomics and Emerging Molecular Therapies of Pancreatic Cancer

Jakub Liu [1], Magdalena Mroczek [2], Anna Mach [3,4], Maria Stępień [5,*], Angelika Aplas [4], Bartosz Pronobis-Szczylik [4], Szymon Bukowski [4], Magda Mielczarek [1,6], Ewelina Gajewska [4], Piotr Topolski [4], Zbigniew J. Król [4], Joanna Szyda [1,6] and Paula Dobosz [4,*]

[1] Biostatistics Group, Wroclaw University of Environmental and Life Sciences, 51-631 Wroclaw, Poland
[2] Centre for Cardiovascular Genetics and Gene Diagnostics, Foundation for People with Rare Diseases, Wagistrasse 25, 8952 Schlieren, Switzerland
[3] Department of Psychiatry, Medical University of Warsaw, 00-665 Warsaw, Poland
[4] Central Clinical Hospital of Ministry of the Interior and Administration in Warsaw, 02-507 Warsaw, Poland
[5] Department of Infectious Diseases, Doctoral School, Medical University of Lublin, 20-059 Lublin, Poland
[6] National Research Institute of Animal Production, Krakowska 1, 32-083 Balice, Poland
* Correspondence: mmaria.stepien@gmail.com (M.S.); paula.dobosz@cskmswia.gov.pl (P.D.)

Simple Summary: This review article is a summary of the current state of knowledge regarding the genetics of pancreatic cancer and a presentation of possible treatment options reflecting genomic medicine advances and a personalised medicine approach. Several oncogenes in which somatic changes lead to the development of tumours, including genes BRCA1/2 and PALB2, TP53, CDKN2A, SMAD4, MLL3, TGFBR2, ARID1A and SF3B1, are involved in pancreatic cancer. Between 4% and 10% of individuals with pancreatic cancer will have a mutation in one of these genes. Six percent of patients with pancreatic cancer have NTRK pathogenic fusion. It is estimated that from 24% to as many as 44% of pancreatic cancers show homologous recombination deficiency (HRD). The most common cause of HRD are inactivating mutations in the genes regulating this DNA repair system, mainly BRCA1 and BRCA2, but also PALB2, RAD51C and several dozen others.

Abstract: The number of cases of pancreatic cancers in 2019 in Poland was 3852 (approx. 2% of all cancers). The course of the disease is very fast, and the average survival time from the diagnosis is 6 months. Only <2% of patients live for 5 years from the diagnosis, 8% live for 2 years, and almost half live for only about 3 months. A family predisposition to pancreatic cancer occurs in about 10% of cases. Several oncogenes in which somatic changes lead to the development of tumours, including genes BRCA1/2 and PALB2, TP53, CDKN2A, SMAD4, MLL3, TGFBR2, ARID1A and SF3B1, are involved in pancreatic cancer. Between 4% and 10% of individuals with pancreatic cancer will have a mutation in one of these genes. Six percent of patients with pancreatic cancer have NTRK pathogenic fusion. The pathogenesis of pancreatic cancer can in many cases be characterised by homologous recombination deficiency (HRD)—cell inability to effectively repair DNA. It is estimated that from 24% to as many as 44% of pancreatic cancers show HRD. The most common cause of HRD are inactivating mutations in the genes regulating this DNA repair system, mainly BRCA1 and BRCA2, but also PALB2, RAD51C and several dozen others.

Keywords: pancreatic cancer; genetics; genomics; molecular therapies

Citation: Liu, J.; Mroczek, M.; Mach, A.; Stępień, M.; Aplas, A.; Pronobis-Szczylik, B.; Bukowski, S.; Mielczarek, M.; Zwolska, A.; Gajewska, E.; et al. Genetics, Genomics and Emerging Molecular Therapies of Pancreatic Cancer. *Cancers* **2023**, *15*, 779. https://doi.org/10.3390/cancers15030779

Academic Editor: Cosimo Sperti

Received: 22 November 2022
Revised: 14 January 2023
Accepted: 18 January 2023
Published: 27 January 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

Pobrano z / Downloaded from Repository of Wrocaw University of Environmental and Life Sciences 2025-10-15 21:43

Pobrano z / Downloaded from Repository of Wrocaw University of Environmental and Life Sciences 2025-10-15 21:43

1. Introduction

Cancer remains an escalating problem worldwide. The reflection of this disturbing trend can be also noticed in Poland and backed up by the statistical data. The number of cases of pancreatic cancers (PC) in 2019 was 3852 (approx. 2% of all cancers) [1]. In 2019, 100,324 people died from cancer in Poland, including 5068 people from pancreatic cancer [1]. The course of the disease is very fast, and the average survival time from the diagnosis is 6 months [2]. Only <2% of patients live for 5 years from the diagnosis, 8% live for 2 years, and almost half live for only about 3 months [2,3]. In the case of resectable tumours, the 5-year survival rate is approximately 20%, but only 10–20% of patients are diagnosed at such a stage of disease that tumour resection is a possible option [4]. Most cases (approx. 90%) of pancreatic cancer occur after the age of 50. The risk of pancreatic cancer increases with age, reaching its peak in the ninth percentile in both sexes (45–50/100,000 inhabitants) [5]. Cases <40 years of age are extremely rare, although globally, there is an increase in this age group—which is usually associated with worse prognosis [2]. Existing studies indicate that the progression of the disease from the initiation to the formation of metastatic pancreatic cancer takes up to 15 years [6]. This creates enough time to recognize the disease early. However, the diagnosis of pancreatic cancer at an early stage, despite recent advances in diagnostic methods, is still very difficult, resulting in the delayed implementation of a potentially effective treatment. Pancreatic cancer patient survival has only slightly improved over the past two decades [7]. At the time of diagnosis, up to 80% of patients already have advanced, metastatic disease, which precludes the effective use of surgical treatment and many other modern therapeutic methods [2].

There are two main types of pancreatic cancer: non-endocrine pancreatic cancer and neuroendocrine pancreatic cancer [8]. The non-endocrine type is divided into four subtypes. The most common of these is adenocarcinoma, which is the cause of roughly 90% of pancreatic cancer diagnoses. This type of cancer emerges in the lining of the pancreatic ducts. Rarely, it can also have its origin in the cells that produce the pancreatic enzymes. Another subtype of the non-endocrine cancer type is colloid carcinoma, which accounts for 1–3% of non-endocrine pancreatic cancers [8]. This subtype has its origin in the intraductal papillary mucinous neoplasm. Colloid carcinoma is easier to treat than other pancreatic cancer types [9].

Adenosquamous carcinoma is the next type of non-endocrine pancreatic cancer. It is the most aggressive type and accounts for 1–4% of diagnoses. The rarest type of pancreatic cancer is squamous cell carcinoma, which is also the least studied subtype, but studies suggest that it has its origins in the pancreatic ducts. The second large class of pancreatic cancers are neuroendocrine pancreatic cancers. This type is rarer than the non-endocrine class, making up less than 5% of cases. This class of cancers develops from the pancreatic gland that releases pancreatic hormones (glucagon and insulin) into the bloodstream [8,9].

On the other hand, the possibility of targeted treatment is the major motivation for the identification of genetic biomarkers in pancreatic cancers. Several genetic biomarkers have proven to be associated with a treatment outcome, among them variants in the following genes: *K-ras*, *hENT1*, *DCK*, *CDA*, *DPD*, *HMLH1*, *TP1DPD*, *HER2* and *SMAD4* [10]. Treatment options depend on the molecular profile of the tumour. For example, tumours with *NTRK* fusion may be treated with a larotrectinib; olaparib treatment targets a mutation in the *BRCA1* or *BRCA2* gene, and pembrolizumab is approved by FDA only in the cancers with MSI-H (microsatellite instability high) or MMR-D (mismatch repair deficiency).

Most of the clinical trials that involve molecular profiles of the pancreas tumour are those including patients with germline *BRCA1*/*BRCA2* pathogenic variants. Most of these involve PARPi, either as a single treatment, together with other oncostatics or as a maintenance therapy. However, in an early stage, there are promising therapies targeting other molecular pathways. For example, there are ongoing trials to implement RAS-directed therapies in pancreatic cancer treatment. One of such implementations is neoantigen T-cell receptor gene therapy, in which a single infusion of genetically engineered autologous

Pobrano z / Downloaded from Repository of Wrocaw University of Environmental and Life Sciences 2025-10-15 21:43

T-cells that target mutant KRAS G12D driver mutation resulted in the regression of metastatic pancreatic cancer [11]. Some studies including immunotherapies, e.g., PD-1 immune checkpoints, are recruiting [12,13]. Though the search for a treatment is ongoing, there are some prominent possibilities for early detection. One of these is the correlation between new-onset diabetes and pancreatic cancer [14]. In a situation in which time is of the essence, this might be a new area to expand upon in early detection. Moreover, it has been shown that patients with diabetes mellitus (DM) have significantly higher median overall survival than those without DM [15].

Meanwhile, in many other cancers, patient mortality is systematically decreasing, which is related, inter alia, to the personalization of therapies and the use of targeted drugs.

2. Genetics of Pancreatic Cancer

Pancreatic cancer, like all cancers, is a genomic disease caused by abnormalities in the sequence or structure of the genomic DNA. Genetic changes leading to neoplasm can be initiated by family predispositions (germline mutations) or by acquired changes (somatic mutations), which occur, for example, because of chronic inflammation in the pancreas [6]. Due to the inaccessible location of many tumours, difficult sampling and the extremely short survival time of patients, the exact molecular basis of the formation of these neoplasms has not been well understood. It is estimated that a family predisposition to pancreatic cancer occurs in about 10% of cases [16,17]. There are also other genetic pathologies that may indirectly lead to the development of pancreatic neoplasms, including hereditary pancreatitis, during which recurring episodes of acute inflammation occur in childhood. There are also several other oncogenes in which somatic changes lead to the development of pancreatic tumours, including genes *BRCA1*, *BRCA2*, *PALB2*, *TP53*, *CDKN2A*, *SMAD4*, *MLL3*, *TGFBR2*, *ARID1A* and *SF3B1* [18–20]. There are also gene fusions, especially in the case of *NTRK*: it is estimated as many as 6% of patients with pancreatic cancer have *NTRK* pathogenic fusion [21]. Currently, the most promising targeted therapies for pancreatic cancer target tumours that are characterised by homologous recombination deficiency (HRD). The mechanism of homologous recombination is responsible for the repair of DNA double-strand breaks [22]. This damage leads to numerous genomic rearrangements and the so-called genomic instability [23]. It is estimated that from 24% to as many as 44% of pancreatic cancers show HRD [23]. The most common cause of HRD are inactivating mutations in the genes regulating this DNA repair system, mainly *BRCA1* and *BRCA2*, but also *PALB2*, *RAD51C* and several dozen others [22].

3. Genomics of Pancreatic Cancer

The mutation's diversity can be studied thanks to recent advances in sequencing technologies (next-generation sequencing; NGS) and computational biology, which have provided the possibility of the reliable detection of molecular alterations, not only for single genes but on the whole genome scale [24]. According to Felsenstein et al., recent genomic studies, including whole exome and whole genome sequencing, have contributed to a better understanding of the complex genomic changes due to a combination of large-scaled structural variants (SV) and widespread single-nucleotide variants (SNV) [25]. As an example, SVs can affect tumour suppressors, including *SMAD4* and *CDKN2A*, by causing homozygous deletion or inactivating alterations [26]. Pancreatic ductal adenocarcinoma subtypes (stable, locally rearranged, scattered, or unstable) have been defined based on the associated SV's features, such as the SVs' count, the predominance of specific SV types and SV's distribution across the genome of each patient [19]. Moreover, *BRCA1*, *BRCA2*, *PALB2* and other genes from the Fanconi anaemia (FA) pathway altered by structural germline or somatic mutations are associated with the unstable subtype, which in turn is caused by the inability to repair double-stranded DNA breaks [27]. In addition, Zhang et al. detected recurrent somatic mutations in multiple genes, including *KRAS*, *TP53*, *CDKN2A*, *SMAD4*, *ARID1A* and *CDKN2B,* in 1080 Chinese patients with pancreatic ductal adenocarcinoma, demonstrating the genomic characteristics on a large scale [28].

Several SNVs have been found to be associated with pancreatic cancer. Some of the most common mutations are the so-called BRCA1 and BRCA2 mutations (as shown in Table 1). In Table 1, the polymorphisms in the PALB2 gene are shown, which have also been found to be associated. Similarly, the mutations in the CDKN2A, KRAS and genes that are related to PC are summarised in Table 1, as well as mutations in the GNAS, ATM, MLH1, MSH2, MSH6 and PSM2 genes (found to cause pancreatic cancer directly or indirectly). Mutations in the genes encoding the CPA1 and CPB1 carboxypeptidases have also been found to be causal (shown in Table 1).

Table 1. Mutations in genes related to pancreatic cancer.

SNP ID or SNP Location	Germline/Somatic	References
SNPs located in the BRCA1 (location: 17q21.31) and BRCA2 gene (location: 13q13.1)	germline	[29]
Mutation in the PALB2 (parner and localiser of BRCA2) gene (location: 16p12.2)	germline	[30]
The KRAS G12D, KRAS G12V and KRAS G12R mutations in the KRAS gene (location: 12p12.1)	somatic	[31]
The KRAS G12R mutation in the KRAS gene (location: 12p12.1)	somatic	[32]
Mutations in the gene (location: 11q22.3) encoding the ATM serine/threonine kinase and the gene (location: 9p21.3) encoding the CDKN2A cyclin-dependent kinase inhibitor 2A	germline	[33]
Mutation in the GNAS gene (chromosome 20)	somatic	[34]
Mutation in the ATM gene	germline	[35]
Mutations in the MLH1, MSH2, MSH6 and PSM2 genes	germline	[36]
Mutations in the genes encoding the CPB1 and CPA1 carboxypeptidases	germline	[37]

Currently, one of the biggest issues in pancreatic cancer treatment is the considerable genetic heterogeneity among pancreatic cancer types and patients. Consequently, many generalised therapies targeting only selected mutations or pathways that are present in a subset of carcinomas are often not effective. Therefore, genome-wide studies are crucial for understanding the contribution of variable genetic mutations to the carcinogenesis, progression and metastasis of pancreatic cancer in order to offer a personalised medicine approach by providing implications for further clinical achievements and drug development [21,38]. Unfortunately, the molecular characterisation in pancreatic cancer is not yet standard in clinical care, but great efforts are being made to use multiple insights provided by recent NGS-based studies in clinical trials [39]. According to Felsenstein et al., genetic alterations may be used to develop diagnostic markers for early detection, disease progression and novel targeted therapies [25]. As an example, modern screening methods, such as the analysis of pancreatic cysts and duodenal fluid, as well as circulating tumour DNA (ctDNA), incorporate the detection of molecular alterations. Mutations in KRAS or GNAS, which are present in more than 96% of patients with a mucin-producing

Pobrano z / Downloaded from Repository of Wrocaw University of Environmental and Life Sciences 2025-10-15 21:43

premalignant neoplastic cyst, can distinguish harmful precursor lesions from benign cysts (e.g., serous cystadenomas) [40,41]. Moreover, altered late driver genes such as TP53 and SMAD4 are even more accurate markers that differentiate cysts requiring urgent clinical intervention from those that can be safely monitored. Mutations in TP53 and SMAD4 are common events in duodenal fluids from patients with pancreatic ductal adenocarcinoma and may be used to differentiate them from control patients. This is possible thanks to the innovative digital NGS technique detecting even low-abundance mutations down to 0.1–1% mutation prevalence. According to Yu et al. [42], mutations of TP53 and SMAD4 genes are often detected in duodenal fluids from patients with pancreatic ductal adenocarcinoma. Interestingly, these mutations allow for distinguishing patients with pancreatic ductal adenocarcinoma from control patients with non-suspicious pancreata. The "liquid biopsy" relying on ctDNA sequencing collected from peripheral blood samples is a powerful technique in metastatic pancreatic cancer recognition that can be used to identify invasive cancers earlier and to monitor patients with recognized cancer. It may be especially useful in the prediction of cancer recurrence, even before a computed tomography scan, and is correlated with survival among metastatic and resected patients [43,44].

It cannot be denied that in the last years, the oncological approach has also been revolutionised by a growing personalisation of the treatment. According to Melisi et al., this precision oncology approach is possible thanks to NGS, which enables genomic profiling [45]. The authors identified genomic alterations in 68 pancreaticoduodenal cancer patients who failed standard treatments. The mutations were characterised based on the ESMO Scale of Clinical Actionability for molecular Targets (ESCAT), which classifies cancer patients who are likely to respond to precision medicines. According to ESCAT, at least one alteration ranking of tier I, II, III or IV was detected in 8, 1, 9 and 12 patients, respectively (44.1%). The most frequent alterations with clinical evidence of actionability (ESCAT tier I-III) affect genes of the RAF (10.3%), BRCA (5.9%) or FGFR pathways (5.9%). The results obtained in this study enabled the selection of patients receiving a molecularly targeted matched therapy. Inconveniently, this genomic profiling method may be used only when standard treatments fail and in young patients with a good performance status [39,45].

The significance of genomic studies is not to be overestimated, and the full characteristics of the pancreatic genome may have promising implications for further clinical significance and drug development. Therefore, not only do known genes need to be studied, but also novel mutated coding regions, as well as non-protein coding genomics locations, such as non-coding RNA [46] and regulatory regions [47], that are likely to have functional significance in carcinogenesis. Such mutations located in functionally important genomic structures may represent putative predictive markers at the early stage of disease, before the onset of symptoms has an essential role in treatment and survival rates.

4. Risk of Metastasis and Genetic Background of the Process

Pancreatic cancer is an aggressive tumour with frequently occurring distant metastases that can appear in virtually every organ. The liver is usually the most affected, followed by the lungs, peritoneum and bones. Various factors are associated with a metastatic pancreas tumour in comparison to an early-stage pancreas tumour, among them non-coding RNAs, transcription factors, growth factors and oxygen conditions as well as mutations [24,48]).

Several studies have aimed to characterise genomic instability in metastatic pancreatic cancer. Campbell et al. reported that an instability pattern in pancreatic metastatic cancer is different, e.g., from that in metastatic breast cancer [49]. The pancreatic cancer could be characterised by breakage–fusion–bridge cycles that predicate specific abnormalities of cell-cycle control, namely, dysregulation of the G1-to-S transition and an intact G2–M checkpoint [49].

Moreover, selected genes play a role in the increase in pancreatic cancer metastatic activity. For example, a *DPC4* expression metastatic pancreatic cancer was higher

Pobrano z / Downloaded from Repository of Wrocaw University of Environmental and Life Sciences 2025-10-15 21:43

compared with early-stage cancers or pancreatic intraepithelial neoplasia. Loss of the DPC4 protein has been shown to correlate with a shortened survival in patients with resectable pancreatic cancers. Among matched primary and metastatic cancer tissues, the loss of MKK4 in immunolabeling was observed in 11% of primary cancers, but in as many as 37% of distant metastases, indicating that MKK4 loss correlates with metastasis formation. On the other hand, other proteins' expression, such as p53, showed no correlation between metastases and the primary tumour in immunostaining.

The differential gene expression in pancreatic cancer has been assessed with different methods, with the global genetic expression being most often assessed with an array expression. However, most of the studies compared a differential genetic expression between tumour and normal tissues, not assessing the genetic and molecular differences between early and metastatic pancreatic tumours [50]. One of the few studies comparing early cancer and metastases concluded that metastatic spread is not accompanied by reproducible changes in gene expression [51]. The studies on mice reported that 25 significant up- and 181 downregulated genes were differentially expressed when comparing the liver metastases with the primary tumour. That shows that on the expression level, some differences between primary and metastatic tumours can be delineated. Eight genes (*PAI-1*, *BNIP3l*, *VEGF*, *NSE*, *RGS4*, *HSP27*, *GADD45A*, *PTPN14*) were additionally validated in a semi-quantitative immunohistochemical analysis and revealed a positive correlation to the array data [52]. Further studies are needed, although they are difficult to perform, as pancreatic tumours are usually already advanced at the time of diagnosis.

Until now, a definite "metastatic cancer signature" could not be identified, although there are events that are generally associated with the metastatic potential of tumours. [49,53]. Interestingly, in pancreatic metastasis, the clonal diversity differs between metastatic sites, with metastases in the lung and liver being monoclonal and those in the peritoneum and diaphragm being polyclonal. These findings indicate that clonal diversity depends on the metastatic site [54] and suggest the role of other mechanisms, such as the microenvironment, in the malignity of pancreatic cancers [53].

5. Hereditary Pancreatic Cancer

Approximately 10–15% of all pancreatic cancers, both exocrine and neuroendocrine, are inherited. Between 10–20% of all mutations and related genetic syndromes, described in the next paragraph, can increase the possibility of developing pancreatic cancer.

Exocrine pancreatic cancer develops from exocrine cells that make digestive enzymes and secrete them into the small intestine. Between 90–95% of all pancreatic cancers are exocrine. Neuroendocrine pancreatic cancer arises from clusters of cells in the pancreas, called islets, that produce hormones. They are neoplasms that exhibit neuroendocrine phenotypes, such as the production of certain neuropeptides, large dense-core secretory vesicles and a lack of neural structures [55]. Hereditary pancreatic cancer can be divided into two categories: FPC (familial pancreatic cancer) and syndromes that do not meet the criteria to be classified as an FPC but contain mutations that may increase the risk of pancreatic cancer. FPC is a ductal adenocarcinoma pancreatic cancer that occurs in families. Patients can be identified with FPC when one or more first-degree relatives have ductal adenocarcinoma. If one first-degree family member is diagnosed with pancreatic cancer, the risk for other family-members is 4–6%. If two family members are diagnosed, the risk increases to 8–12%, and with over two positive family members, the risk is 17–32% [56].

The association of familial pancreatic cancer with other malignant neoplasms was also investigated using the "Swedish Family-Cancer Database", which includes over 11.5 million individuals. The results of this study demonstrated an association between pancreatic cancer in parents and melanoma in offspring. An increased risk of cancers of the lung, breast, small intestine, colon, testes and cervix has also been observed in people with a family history of pancreatic cancer [57].

Pobrano z / Downloaded from Repository of Wrocaw University of Environmental and Life Sciences 2025-10-15 21:43

Pobrano z / Downloaded from Repository of Wrocaw University of Environmental and Life Sciences 2025-10-15 21:43

6. Pancreatic Cancer Associated with Cancer Syndrome

Most of the scientific reports indicate an association between the appearance of specific variants, genes or environmental factors as risk factors for pancreatic cancer. However, few of them show pancreatic cancer as a risk factor for other types of cancer.

In their study, Cote and al. assessed whether a family history of pancreas cancer is a risk factor for first-degree relatives for ten specific cancers, i.e., breast, colon, bladder, lungs, lymphoma, melanoma, ovary, pancreas, prostate head, and neck. A doubled risk of lymphoma and ovarian cancer was observed in first-degree relatives of individuals with pancreatic cancer [58].

Pancreatic cancer appears to be a genetically heterogeneous disease that is an integral alteration of many cancer syndromes. Several cancer syndromes predispose one to an increased risk of pancreatic cancer. These include hereditary pancreatitis (HP), hereditary non polyposis colorectal cancer (HNPCC) or Lynch syndrome (LS), hereditary breast and ovarian cancer (HBOC), familial atypical multiple mole melanoma syndrome (FAMMM), Peutz–Jeghers syndrome (PJS), Ataxia–telangiectasia (AT), familial adenomatous polyposis (FAP), familial pancreatic cancer, familiar breast cancer, von Hippel–Lindau syndrome (VHL), Li–Fraumeni syndrome (LFS) and Fanconi anaemia [59–69]. An additional group represents cancer syndromes associated with pancreatic neuroendocrine tumours (pNETs). Approximately 1% to 2 % of pancreatic neuroendocrine tumours are present in the context of hereditary disorders [70]. These include Multiple Endocrine Neoplasia type 1 (MEN1) (Wermer syndrome), von Hippel–Lindau disease, neurofibromatosis 1 (NF-1) and the tuberous sclerosis complex (TSC) [71]. Based on this, several potential candidate genes for PC have been proposed. In a study by Hu et al., a significant association was observed between pancreatic cancer and inherited germinal mutations in the *CDKN2A*, *TP53*, *MLH1*, *BRCA2*, *ATM* and *BRCA1* genes [72]. Further studies have identified other deleterious mutations in pancreatic cancer patients in genes *CHEK2*, *MUTYH/MYH*, *BARD1*, *MSH2*, *NBN*, *PALB2* and *PMS2* [73]. The best-known syndromes associated with PC are hereditary breast and ovarian cancer syndrome and familial atypical nevus and melanoma syndrome [73]. Mutations in the *BRCA1*, *BRCA2* and *CDKN2A* genes were found in 7.4% of PC patients [74]. The risk of PC in cancer syndromes is summarised in Table 2.

Table 2. Risk of development of pancreatic cancer.

Cancer Syndrome	Gene(s)	Location	Mode of Inheritance	Relative Risk for PC Compared with the General Population	Cumulative Risk of PC (%)	References
HBOC	*BRCA2*	13q13.1	AD	3.5–10 fold	3–5% by age 70 2–7% lifetime risk	[60,75,76]
	BRCA1	17q21.31	AD	2.26–3 fold	1–3% lifetime risk	[75–77]
	PALB2	16p12.2	AD	2.37 fold	2–3% up to 80 years	[62]
PJS	*STK11/LKB1*	19p13.3	AD	76–132 fold	11–36% to age 65–70	[61,78,79]
FAMMM	*CDKN2A*	9p21.3	AD	13–38 fold	17% by age 70 up to 20% by the age of 75	[63,80,81]
VHL	*VHL*	16p13	AD	-	5–17%	[64,82]
LFS	*TP53*	17p13	AD	6–7.3 fold	-	[65,83]
FA	Fanconi anaemia complex genes: *FANCN (PALB2)* *FANNG* *FANCM* *FANCL* *FANCD1(BRCA2)* *FANCC*	16p12.2 9p13.3 14q21.2 2p16.1 13q13.1 9q22.32	AR	-	-	[66,84]

HP	*PRSS1* *SPINK1*	7q35 5q32	AD	35–70 fold	7.2–18% lifetime risk	[67,85]
FAP	*APC*	5q21	AD	4.46 fold	2% lifetime risk	[69,86,87]
LS (HNPCC)	*MLH1,* *MSH2,* *MSH6,* *PMS2,* *EPCAM*	3p22.2 2p21 2p16.3 7p22.1 2p21	AD	5–8.6 fold	1.1% up to age 50 3.9% up to age 70 6.2% up to age 75	[68,88,89]
AT	*ATM*	11q22.3	AR	5.7–6.5 fold	0.08% by age 30 6.3% by age 70 9.5% by age 80	[71]
MEN1	*MEN1*	11q13	AD	-	Frequency PET: 80–100% (microscopic) 20–80% (clinical)	[71]
NF-1	*NF-1*	17q11.2	AD	-	Frequency PET: Uncommon (0–10%)	[71]
TSC	*TSC1* *TSC2*	9q34 16p13	AD	-	Frequency PET: Uncommon	[71]

AD—autosomal dominant; *APC*—adenomatous polyposis coli gene; AR—autosomal recessive; AT—ataxia–telangiectasia; *ATM*—ataxia–telangiectasia mutated gene, *BRCA1*—breast cancer 2 gene; *BRCA2*—breast cancer 2 gene; *CDKN2A*—cyclin-dependent kinase inhibitor 2 A gene; *EPCAM*—epithelial cell adhesion molecule gene; FA—Fanconi anemia, FAMMM—familial atypical multiple mole melanoma syndrome; FAP—familial adenomatous polyposis; HBOC—hereditary breast and ovarian cancer; HNPCC—hereditary non-polyposis colorectal cancer; HP—hereditary pancreatitis; LFS—Li–Fraumeni syndrome; LS—Lynch syndrome; MEN1—multiple endocrine neoplasia type 1; *MLH1*—MutL homolog 1 gene; *MSH2*—MutS protein homolog 2 gene; *MSH6*—MutS homolog 6 gene; NF-1—neurofibromatosis1 (von Recklinghausen's disease); *PALB2*—partner and localiser of BRCA2 gene; PC—pancreatic cancer; PET—pancreatic endocrine tumour; PJS—Peutz–Jeghers syndrome; *PMS2*—PMS1 homolog 2, mismatch repair system component 2 gene; *PRSS1*—protease, Ser1 gene; *SPINK1*—serine peptidase inhibitor Kazal type 1 gene; *STK11*—serine/threonine kinase 11 gene; *TP53*—tumour protein 53 gene; TSC—tuberous sclerosis complex; *TSC1*—tuberous sclerosis complex 1 gene; *TSC2*—tuberous sclerosis complex 2 gene; VHL—von Hippel–Lindau syndrome; (-)—no data available.

6.1. Hereditary Breast and Ovarian Cancer

Hereditary breast and ovarian cancer is inherited as an autosomal dominant disorder characterised by increased risk of breast cancer, ovarian cancer and other cancers such as prostate cancer, melanoma and pancreatic cancer. The disease is associated with pathogenic or likely pathogenic variants in one of the two genes *BRCA1* or *BRCA2*. Their protein products are involved in DNA repair, recombination, transcription and tumour suppression [90]. Previous studies reported an increased risk of PC in both *BRCA1* (1–3%) [77,91] and *BRCA2* (2–7%) carriers [92,93] (Table 2). In particular, the risk appears to be evident in families of Ashkenazi Jewish descent, where *BRCA1/2* founder mutations were identified in 5.5% of Ashkenazi patients with PC [94]. Interestingly, in most studies, the risk of MS in BRCA2 carriers was higher than in *BRCA1* carriers [75]. Recent studies have also revealed genes other than BRC1/2 that have been identified as a causative factor in HBOC syndrome. These are mainly genes associated with an increased risk of breast and ovarian cancer, but certain genes also have a proven link to pancreatic cancer. These include the genes *PALB2, TP53*, *STK11, ATM, MLH1* and *MSH2*. Interestingly, they are also strongly associated with other cancer syndromes [95].

6.2. Familial Atypical Mole and Multiple Melanoma Syndrome

Familial atypical multiple mole melanoma syndrome is an autosomal dominant inherited disorder characterised by multiple dysplastic nevi and melanoma [96]. Approximately 20–40% of patients with FAMMM have a germinal mutation in the *CDKN2A* gene, located on chromosome 9p21, but only about 1% have a mutation in another *CDK4* gene on chromosome 12q14 [97,98]. Some *CDKN2A* mutations have also been linked to

Pobrano z / Downloaded from Repository of Wrocaw University of Environmental and Life Sciences 2025-10-15 21:43

malignancies other than FAMMM, particularly PC [97]. The association between FAMMM syndrome and PC is well documented. *CDNK2* mutations are associated with an increased pancreatic cancer incidence of up to 20% by age 75 [99]. The risk of PC in patients with FAMMM is estimated to be 13–22 times higher than in the general population and additionally increases to 38-fold in FAAMM subjects with *CDKN2A* mutations [100,101]. Interestingly, nearly 90% of sporadic PDAC harbour alterations in CDKN2 [49].

6.3. Lynch Syndrome

Lynch syndrome, also known as hereditary non polyposis colorectal cancer, is an autosomal dominant disease associated with an increased predisposition to colorectal, gastric and endometrial cancers [102]. LS is caused by germline mutations in mismatch repair genes (MMR), including genes encoding DNA reparation proteins: *MLH1*, *MSH2*, *MSH6*, *PMS2* and *EPCAM* gene, which lead to epigenetic silencing of *MSH2* [103]. The malfunctioning variants are referred to as *path_MLH1*, *path_MSH2*, *path_MSH6* and *path_PMS2* [104]. Mutations in the *MLH1* and *MSH2* genes have a greater impact on DNA repair; therefore, patients carrying these mutations have a significantly higher risk of developing cancer compared to patients with *MSH6* or *PMS2* mutations [103,105]. Kastrinos et al. demonstrated that patients from families with germline *MMR* gene mutations have a cumulative risk of pancreatic cancer ranging from 1.31% up to the age of 50 years to 3.68% up to age of 70 years [68]. This represents an 8.6-fold increase compared with the general population [68]. Similar results of the cumulative incidence of PC for *path_MMR* variants carriers were reported by Møller et al., who estimated a cumulative risk of 1.1% up to age 50, 3.9% up to 70 and 6.2% up to 75 years of age [88].

6.4. Peutz–Jeghers Syndrome

Peutz–Jeghers syndrome is a rare autosomal dominant disorder. PJS is most often caused by a germline mutation in the serine/threonine kinase 11 or liver kinase B1 STK11/LKB1 genes on chromosome 19p13.3 [106]. The *STK11/LKB1* is a tumour suppressor gene that plays an important role in regulating the cell cycle [107]. Mutations in *STK11* are identified in 50% to 80% of families with PJS. In other patients, PJS is probably the cause of de novo mutations [108]. It is characterised by hamartomatous gastrointestinal polyps and pigmented cutaneous and mucocutaneous macules [108]. Individuals with Peutz–Jeghers syndrome have an increased risk for developing colorectal, gastric system, breast, uterine, cervical, lung, ovarian, testicular and pancreatic cancers [109]. The literature has reported that the cumulative risk of developing PC ranges from 11 to 36% at the age of 70 years, which presents a 136-fold increase compared with the general population [61,78].

6.5. Li–Fraumeni Syndrome

Li–Fraumeni syndrome is a rare hereditary autosomal dominant disorder due to germline pathogenic or likely pathogenic variants in the *TP53* gene. Approximately 7% to 20% of patients with germline *TP53* pathogenic variants have mutations derived de novo. LFS is characterised by the development of a wide spectrum of childhood and adult-onset malignancies. The main cancers are adrenocortical carcinomas, breast cancer, central nervous system tumours, osteosarcomas and soft-tissue sarcomas [110]. The lifetime risk of cancer in individuals with LFS approaches 75% in males and almost 100% in females [110,111]. LFS is also associated with an increased risk of several additional cancers, including leukaemia, lymphoma, gastrointestinal cancers, cancers of head and neck, kidney, skin, lung, larynx, ovary, thyroid, prostate, testis and pancreas. In addition to these cancers, patients with LFS have a 7.3-fold increased risk of PDAC [65].

Pobrano z / Downloaded from Repository of Wrocaw University of Environmental and Life Sciences 2025-10-15 21:43

6.6. Familial Adenomatous Polyposis

Familial adenomatous polyposis syndrome is an autosomal dominant disease, predominantly caused by a germline mutation in the adenomatous polyposis coli *(APC*) gene on chromosome 5q21 [112]. It is a disease of high penetrance and variable expression [86]. Approximately 20–30% of FAP cases arise from de novo mutations, without a family history [113]. The mutation site in the *APC* gene determines the progression, clinical phenotypic variations (including Gardner and Turcot syndromes) and the risk of development of extracolonic manifestations of the disease [112,114]. FAP is characterised by hundreds to thousands of adenomatous polyps of the gastrointestinal mucosa [115]. FAP is also correlated with the incidence of other malignancies, mainly colorectal cancer, gastric, duodenal, hepatoblastoma and desmoid tumours [116]. The lifetime incidence of pancreatic cancer in patients with FAP is relatively low. The lifetime risk has been estimated at around 2%, but it is still four times higher than in the general population [69,87]. Furthermore, reports on the genetic link between the two disorders are limited and remain inconclusive. Hori et al. demonstrated sporadic *APC* mutations in pancreatic cancer [117], which has not been confirmed in other studies [118]. Another group of investigators hypothesised that *APC* mutation may play a role in the initiation of pancreatic cancer [119].

6.7. Ataxia–Telangiectasia

Ataxia–telangiectasia is an autosomal recessive disorder characterised by neurological symptoms, such as cerebellar atrophy with progressive ataxia, and a marked predisposition to cancer [120]. Ataxia telangiectasia is due to mutations of the ATM gene, whose product is involved in DNA double-break repair [121]. In a study of 166 familial pancreatic cancer probands, 2.4% were carriers of the deleterious *ATM* variant [120]. Large-scale sequencing-based studies have identified up to 18% of *ATM* mutations in some cohorts of patients with pancreatic ductal adenocarcinoma [122–124]. The risk of pancreatic cancer among *ATM* pathogenic variant carriers has been estimated at 6.3% by age 70 (Table 2) [125].

6.8. Fanconi Anaemia

Fanconi anaemia is characterised by bone marrow failure, developmental abnormalities and an increased risk to develop malignancies. FA is a predominantly autosomal recessive disease. Nineteen FA-related genes (FANC) whose products suppress interstrand crosslink sensitivity have been described. FA patients harbour biallelic mutations in a particular FA gene, with the exceptions of *FANCB (*X-linked) and *FANCR*/*RAD51,* which are autosomal dominant [126]. Mutations in most FA genes lead to a chromosomal instability disorder. Several studies have reported FA gene mutations in pancreatic cancers [127–130]. Roberts et al. reported protein-truncating variants of AF (*BRCA2, PALB2, FANC, FANCG* and *FANCM*) in individuals with familial pancreatic cancer [84]. In another study, Rogers et al. showed that germline and somatic mutations in *FANCC* and *FANCG* may contribute to pancreatic cancer [128]. Mutations in the *FANCC* gene have been detected in two young-onset pancreatic cases, but none in *FANCG* [131].

6.9. Multiple Endocrine Neoplasia Type 1

Multiple endocrine neoplasia type 1 is a rare autosomal dominant and hereditary endocrine tumour syndrome with high penetrance. Familial MEN1is characterised by the combined occurrence of tumours of the parathyroid glands, endocrine pancreas, duodenum, anterior pituitary and, less commonly, of the stomach, adrenal gland, thymus and lungs [132]. MEN1 is caused by several germline mutations in the tumour suppressor *MEN1* gene that encodes the protein menin [133]. In a study by Ito et al. in patients with pNETs, *MEN-1* was detected in 10% of patients [134]. On the other hand, pancreatic endocrine tumours occur in 40–70% of MEN-1 patients [135]. MEN-1-associated pancreatic

Pobrano z / Downloaded from Repository of Wrocaw University of Environmental and Life Sciences 2025-10-15 21:43

Pobrano z / Downloaded from Repository of Wrocaw University of Environmental and Life Sciences 2025-10-15 21:43

endocrine tumours are most commonly gastrinomas and non-functioning tumours [136]. Pancreatic endocrine tumours are the leading cause of death in MEN-1 patients [137].

6.10. Von Hippel–Lindau Syndrome

Von Hippel–Lindau syndrome is a hereditary autosomal dominant disease caused by mutations in the *VHL* gene located on chromosome 3. It is a disease with high penetration, early onset and a high prevalence of clinical symptoms. VHL is characterised by multi-organ tumours of the central nervous system, kidney, adrenal and reproductive organs and pancreas. The most common type of tumour in VHL is hemangioblastoma. Renal cell carcinoma (RCC) and endolymphatic and pancreatic neuroendocrine tumours are also diagnosed in individuals with VHL [64,138]. Approximately 35–70% of patients with VHL have pancreatic lesions, and 5–17% develop neuroendocrine tumours [82,139]. In a study by Hammel et al., pancreatic disorders were found in 77% of patients with VHL. This included true cysts (91.1%), serous cystadenomas (12.3%), neuroendocrine tumours (12.3%) or combined lesions (11.5%). Interestingly, in 7.6% of patients, the pancreas was the only organ affected [140]. Reich et al. reported an age-dependent cumulative probability of pancreatic neuroendocrine neoplasm of 10 % up to 40 and 26% up to 60 years of age [141].

6.11. Neurofibromatosis 1

Neurofibromatosis 1, also known as von Recklinghausen disease, is an autosomal dominant disorder characterised mainly by neurocutaneous manifestations [142]. It is caused by a mutation in the *NF-1* gene, which encodes the neurofibromin protein, leading to uncontrolled cell proliferation and the development of some tumours [143]. Gastrointestinal involvement is rare but is associated with a significant risk of malignancy. To date, there are only a few case reports linking NF-1 to pancreatic neuroendocrine tumours [70,144,145].

6.12. Tuberous Sclerosis (Bourneville's Disease)

Tuberous sclerosis is an autosomal dominant neurocutaneous multisystem disorder. The disease is due to mutations in one of two genes—*TSC1* located on chromosome 9q34 encoding for the protein hamartin, or *TSC2* located on chromosome 16p13.3 encoding for the protein tuberin [146,147]. Among familial cases of TSC, about half show an association with the *TSC1* gene and half with the *TSC2* gene [148]. Pancreatic endocrine tumours have been reported in a small percentage of patients with TSC [149–151].

As presented above, most of the studies provide evidence for the contribution of genetic background in the development of cancer syndromes and related PC. Considering this evidence, it seems reasonable to include individuals with cancer syndrome in screening. The aim of these screenings would be to identify pre-cancerous lesions, which would enable actions to be taken to prevent the development of cancer or to detect cancer at an early, still treatable stage.

7. Targeted Therapy Overview and Future Perspectives

Though immunotherapy, microbiological therapy and targeted therapy (examples in Table 3 below) remain the most promising treatment options for pancreatic cancer at the moment, we should not forget about traditional methods, such as surgery (often combined with neoadjuvant therapy), radiation and chemotherapy (including adjuvant therapy) [152,153]. Further, palliative care plays an important role in many PC cases, as this is still a tumour with a high mortality rate [153]. Chemotherapy of pancreatic cancer depends on the stage and previous/planned treatment; MFOLFIRINOX (modified leucovorin, 5-fluorouracil, irinotecan, oxaliplatin) or gemcitabine with capecitabine are often used for 6 months after radical resection [153]. There is currently no immunotherapy approved for PC treatment, as it is considered to be a less immunogenic cancer; however, it

Pobrano z / Downloaded from Repository of Wrocaw University of Environmental and Life Sciences 2025-10-15 21:43

may be used in combination with chemotherapy, chemoradiotherapy, vaccines and cytokine antagonism [153]. In recent years, research on modern therapies is becoming more and more extensive, and it is probable that new methods will be combined with traditional methods of treating PC [152,153].

When it comes to developing new therapies, there is hope in organoid models that can simulate the characteristics of pancreatic tumours [154]. This is a rapidly evolving preclinical model that allows for studying the characteristics of pancreatic cancer and leads to an improvement in patient outcomes [154].

Currently, the standards for first-line treatment in PDAC are FFX and GnP regimens. Unfortunately, in most cases, first-line therapy is not successful, but there are still not enough RTCs. It is suggested that FFX followed by GnP or GnP followed by FFX is the most feasible or beneficial approach [155]. The high incidence of germline mutations in PDAC patients makes targeted therapy tempting [155].

Given the frequent instability of the genome, which refers to a range of genetic alterations, from point mutations to chromosomal rearrangements caused by damage to the DNA repair mechanisms, various interventions to accumulate double-strand breaks and to induce apoptosis have been investigated in pancreatic tumours, the most effective of which is treating patients with inhibitors of poly-(adenosine diphosphate ribose) polymerase (PARPi for short) [156,157]. The European Medicines Agency registration allows for the administration of PARPi as a maintenance therapy in patients with pancreatic cancer, in whom a germinal mutation in *BRCA1/2* was detected, and disease progression was not observed for a minimum of 16 weeks after first-line treatment with platinum therapy. Olaparib has been also investigated in clinical trials as a monotherapy [29]. Olaparib proved to be an effective maintenance therapy for BRCA germline patients and prolonged progression-free survival [29]. There are numerous indications that it is possible, after advanced genetic testing, to expand the group of patients receiving PARPi [158]. The results of clinical trials in ovarian cancer, in which PARPi have been used for a long time, demonstrated the effectiveness of PARPi also in patients with somatic changes in the *BRCA1/2* genes, as well as with characteristic changes in the genome associated with HRD or referred to as genomic instability. Moreover, the latest registration of PARPi, called Niraparib, assumes that this drug is administered to patients with ovarian tumours, regardless of their mutational status [158].

With the development and rising accessibility of the omics diagnostic methods and data analysis, we can better characterise the genetics of pancreatic cancer, although the relevance of genetic markers and signatures still needs to be understood [159,160]. High-throughput methods have revealed other non-germline genetic variants involved in PDAC pathogenesis. The in silico analysis of expression data has revealed new therapeutic perspectives, such as bleomycin, as potential therapeutic agents for PDAC [161], although other non-DNA biomarkers can characterise PDAC on the molecular level and the response to therapy [162]. Although the NTRK pathogenic variant is rare with PDAC, such tumours can be treated with larotrectinib [163], and first reports on such cases have been published [164].

Due to the short life expectancy of patients and the time-consuming nature of such analyses, there have been few studies to date extending the indication for the administration of PARPi in pancreatic cancer.

Precision medicine is very promising for the treatment and early diagnosis of pancreatic cancer. It is also worth mentioning that there are three types of new-generation treatments for pancreatic cancer-associated mutations: nucleic acid drugs, small molecules and antibodies. Nucleic acid therapy has many advantages over traditional drugs, for example: the inhibition of previously inaccessible targets, faster and longer lasting responses than protein inhibition, less severe side effects and short development duration [165]. Since 90% of pancreatic cancers have a KRAS mutation, it is the most common target. Anti-tumour activity was detected in mice bearing this mutation. It was also tested in more than 1000 clinical trials, and few ASOs showed positive results. In addition, ASO targeting XIAP combined with gemcitabine showed some effect. Fifty-eight percent of patients that received this combination lived 6 months or longer [165].

There is no approved RNAi treatment for pancreatic cancer, although in clinical trials, Golan et al. targeted the most common gene—KRAS. Unfortunately, the clinical trials did not show promising results—in 83% of patients, no change accrued, and in 17%, there was very little effect, but the treatment is safe [165]. It can be implemented in combination with chemotherapy. Nishimura et al. presented a trial in which a fine-needle injection of RNA oligonucleotide-targeted CHST15 turned out to be safe [166]. Although there are many kinds of nucleic acid drugs, such as small interfering RNA, decoys, CpG oligodeoxynucleotides, microRNA, decoys and aptamers, which are very promising, unfortunately, only 30% of patients' mutations can be treated with drugs, because it is very hard to inhibit targeted RNA in humans: *PALB2*, *ATM*, *HER2*, *MET*, *MLH1*, *MSH2*, *MSH6*, *PMS2*, *KRAS*, *BRCA1 and 2*, *PI3CA*, *PTEN*, *CDKN2A*, *BRAF* and *FGFR1* [167].

Table 3. Targeted therapies currently used and in clinical trials among pancreatic cancer patients.

Treatment	Gene and Indication	Side Effects	References
PARPi (e.g., Niraparib, Olaparib)	*BRCA1/2*—known or suspected BRCA gene mutation, no worsening after the treatment with platinum resistant agent; also targeting tumours with HRD, including mutations in: *PALB2, RAD51C, BRCA 1/2*	nausea, vomiting, diarrhoea or constipation, fatigue, anaemia, leukaemia, taste changes	[23,30,157–159]
Erlotinib	*EGFR*—advanced pancreatic cancer may be administered in combination with gemcitabine	neutropenia, febrile neutropenia, fatigue, nausea, infections, vomiting, mucositis	[168–170]
Larotrectinib and entrectinib	*NTRK*—pancreatic cancer with NTRK mutations, resistance to other treatment	fatigue, nausea, vomiting, constipation, weight gain, and diarrhoea	[163,164]
Everolimus	inhibitor of mammalian target of rapamycin (mTOR); mutations in TSC1 and TSC2 genes	stomatitis, infections, diarrhoea, peripheral edema, fatigue rash, anaemia, hypercholesterolemia, lymphopenia, elevated aspartate transaminase (AST), fasting hyperglycemia	[171]
Belzutifan	hypoxia-inducible factor-2 alpha (HIF-2α) inhibitor, used for the treatment of von Hippel–Lindau disease-associated cancers	anaemia, low oxygen levels—shortness of breath, headache, dizziness, tiredness; nausea; increased blood sugar (increased thirst or urination, dry mouth, fruity breath), abnormal kidney function tests	[172–175]
Palbociclib	*CDKN2A mutations*	nausea, diarrhoea, vomiting, decreased appetite, changes in taste, tiredness, numbness or tingling in	[176]

Pobrano z / Downloaded from Repository of Wrocaw University of Environmental and Life Sciences 2025-10-15 21:43

		arms, hands, legs, and feet, sores on the lips, mouth or throat	
[experimental/clinical trials]	*TGFBR2*—known or suspected *TGFBR2* gene mutation or its pathway	unknown	[19–21,177,178]
[experimental/clinical trials]	*SMAD4 mutations*	unknown	[179–181]
[experimental/clinical trials]	*MLL3 mutations*	unknown	[182–185]

8. Drug Resistance Induced by Genetic and Genomic Alterations

Designing effective treatment regimens requires an understanding of all potential mechanisms of therapeutic resistance. Genetic and epigenetic changes in tumour cells may play a significant role in resistance to chemotherapy, radiotherapy and immunotherapy for PC.

Among the genetic alterations contributing to drug resistance, mutations in the tumour suppressor genes *TP53*, *SMAD4* and *CDKN2A,* and especially mutations in the KRAS oncogene, have been highlighted [19,186]. In pancreatic cancer, KRAS is the most frequently mutated oncogene [187]. The activation of mutations in KRAS, observed in more than 90 % of pancreatic ductal adenocarcinoma cases, are a critical factor involved in tumour initiation and progression. Studies have shown a positive correlation of KRAS with gemcitabine resistance [188]. Although the first therapies based on KRAS inhibitors have been introduced, they may only benefit a small group of patients. Sotorasib, adagrasib and other currently tested inhibitors mainly target the *KRAS* G12C mutation, and there are no approved drugs targeting KRAS G12D and G12V [189,190].

In addition, drug resistance may be acquired during therapy and may result from genetic mutations arising during the treatment process as well as adaptive mechanisms and natural clonal selection of tumour cells [191]. Among patients treated with KRAS G12C inhibitors, the resistance mechanism was detected in 45% of patients. Acquired KRAS gene alterations associated with the observed resistance to treatment included G12D/R/V/W, G13D, Q61H, R68S, H95D/Q/R, Y96C and amplifications in *KRAS* [192]. It is thought that secondary mutations may lead to an increase in the active KRAS protein bound to GTP and also prevent binding to the drug [193]. Awad et al. observed resistance to both adagrasib and sotarisib associated with R68S and Y96C mutations, and H95D, H95Q and H95R mutations resulted in resistance to adagrasib only. Interestingly, in patients with acquired resistance to adagrasib, additional pathogenic alterations occurring not in the *KRAS* gene itself, but they were identified in the RTK-RAS signalling pathways. These included mutations in the *NRAS*, *BRAF*, *MAP2K1*, *RET*, *NF1* and *PTEN* genes, and they also played a significant role in the development of treatment resistance [192]. Another aspect is resistance to radiotherapy, which has been found for mutations in *KRAS*, *TP53* and *CDKN2A* [194].

The occurrence of resistance may also result from epigenetic changes, such as the regulation of gene expression by small non-coding RNAs (miRNAs). The results from studies have shown that the expression of specific miRNAs in PC correlates with sensitivity or resistance to chemotherapy [195,196]. Gemcitabine-resistant cells were characterised by low expression levels of miR-101-3p and miRNA-124 [197,198]. Furthermore, gemcitabine treatment led to secondary changes in the levels of individual miRNAs (miRNA-17-5p, miRNA-21, miRNA-203), consequently leading to chemoresistance [199–201].

Pobrano z / Downloaded from Repository of Wrocaw University of Environmental and Life Sciences 2025-10-15 21:43

9. Conclusions

Advances in sequencing technology are now reducing the time needed to obtain results and opening new opportunities for patients. Moreover, advanced genomic testing also provides information on potential susceptibility to other available targeted therapies, giving patients the chance to receive effective treatment. The current state of knowledge regarding the genetics of pancreatic cancer and the emerging new possibilities of available treatment options reflects advances in genomic medicine and personalised medicine, providing new hopes for the near future.

Author Contributions: Conceptualization—Z.J.K., J.S., P.D.; funding acquisition and drafting the manuscript—J.S., J.L., M.M. (Magdalena Mroczek), B.P.-S., S.B., P.D.; investigation—J.L., M.M. (Magda Mielczarek), A.M., M.S., M.M. (Magdalena Mroczek), E.G., P.T.; supervision—J.S., M.M. (Magdalena Mroczek), P.D.; writing—original draft—J.L., M.M. (Magdalena Mroczek), A.A., B.P.-S., S.B., P.T.; writing—review and editing—M.M. (Magda Mielczarek), M.S., M.M. (Magdalena Mroczek), Z.J.K.; final proofreading: Z.J.K., J.S., P.D. All authors have read and agreed to the published version of the manuscript.

Funding: The APC is financed by Wroclaw University of Environmental and Life Sciences.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors have no relevant financial or non-financial interest to disclose.

Pobrano z / Downloaded from Repository of Wrocaw University of Environmental and Life Sciences 2025-10-15 21:43

References

1. Didkowska, J.; Wojciechowska, U.; Olasek, P.; Caetano Dos Santos, F.L.; Michałek, I.M. *Cancer in Poland in 2019*; Ministry of Health, Warsaw, Poland, 2021.
2. Michaud, D.S. Epidemiology of Pancreatic Cancer. *Minerva Chir.* **2004**, *59*, 99–111.
3. Poruk, K.E.; Firpo, M.A.; Adler, D.G.; Mulvihill, S.J. Screening for Pancreatic Cancer. *Ann. Surg.* **2013**, *257*, 17–26. https://doi.org/10.1097/SLA.0b013e31825ffbfb.
4. Lampe, P.; Grabarczyk, A.; Kuśnierz, K. Theoretical and Realistic Aspects in the Diagnosis of Pancreatic Cancer. *Postępy Nauk. Med.* **2015**, *28*, 4–10.
5. Vujasinovic, M.; Dugic, A.; Maisonneuve, P.; Aljic, A.; Berggren, R.; Panic, N.; Valente, R.; Pozzi Mucelli, R.; Waldthaler, A.; Ghorbani, P.; et al. Risk of Developing Pancreatic Cancer in Patients with Chronic Pancreatitis. *J. Clin. Med.* **2020**, *9*, 3720. https://doi.org/10.3390/jcm9113720.
6. Munigala, S.; Subramaniam, D.S.; Subramaniam, D.P.; Burroughs, T.E.; Conwell, D.L.; Sheth, S.G. Incidence and Risk of Pancreatic Cancer in Patients with a New Diagnosis of Chronic Pancreatitis. *Dig. Dis. Sci.* **2022**, *67*, 708–715. https://doi.org/10.1007/s10620-021-06886-7.
7. Khalaf, N.; El-Serag, H.B.; Abrams, H.R.; Thrift, A.P. Burden of Pancreatic Cancer: From Epidemiology to Practice. *Clin. Gastroenterol. Hepatol.* **2021**, *19*, 876–884. https://doi.org/10.1016/j.cgh.2020.02.054.
8. Wolfgang, C.L.; Herman, J.M.; Laheru, D.A.; Klein, A.P.; Erdek, M.A.; Fishman, E.K.; Hruban, R.H. Recent Progress in Pancreatic Cancer. *CA Cancer J. Clin.* **2013**, *63*, 318–348. https://doi.org/10.3322/caac.21190.
9. Johns Hopkins Medicine Pancreatic Cancer Types. Available online: https://www.hopkinsmedicine.org/health/conditions-and-diseases/pancreatic-cancer/pancreatic-cancer-types (accessed on 10 January 2023).
10. Fotopoulos, G. Genetic Factors Affecting Patient Responses to Pancreatic Cancer Treatment. *Ann. Gastroenterol.* **2016**, *29*, 466–476. https://doi.org/10.20524/aog.2016.0056.
11. Leidner, R.; Sanjuan Silva, N.; Huang, H.; Sprott, D.; Zheng, C.; Shih, Y.-P.; Leung, A.; Payne, R.; Sutcliffe, K.; Cramer, J.; et al. Neoantigen T-Cell Receptor Gene Therapy in Pancreatic Cancer. *N. Engl. J. Med.* **2022**, *386*, 2112–2119. https://doi.org/10.1056/NEJMoa2119662.
12. Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins; Merck Sharp & Dohme LLC.; Verastem, Inc. Study of Pembrolizumab with or without Defactinib Following Chemotherapy as a Neoadjuvant and Adjuvant Treatment for Resectable Pancreatic Ductal Adenocarcinoma. Available online: https://www.clinicaltrials.gov/ct2/show/NCT03727880 (accessed on 10 January 2023).

13. Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins; Bristol-Myers Squibb. Trial of Neoadjuvant and Adjuvant Nivolumab and BMS-813160 With or Without GVAX for Locally Advanced Pancreatic Ductal Adenocarcinomas. Available online: https://clinicaltrials.gov/ct2/show/NCT03767582 (accessed on 10 January 2023).
14. Sharma, A.; Kandlakunta, H.; Nagpal, S.J.S.; Feng, Z.; Hoos, W.; Petersen, G.M.; Chari, S.T. Model to Determine Risk of Pancreatic Cancer in Patients With New-Onset Diabetes. *Gastroenterology* **2018**, *155*, 730–739.e3. https://doi.org/10.1053/j.gastro.2018.05.023.
15. Badowska-Kozakiewicz, A.; Fudalej, M.; Kwaśniewska, D.; Durlik, M.; Nasierowska-Guttmejer, A.; Mormul, A.; Włoszek, E.; Czerw, A.; Banaś, T.; Deptała, A. Diabetes Mellitus and Pancreatic Ductal Adenocarcinoma—Prevalence, Clinicopathological Variables, and Clinical Outcomes. *Cancers* **2022**, *14*, 2840. https://doi.org/10.3390/cancers14122840.
16. PanCan: Genetic Hereditary Pancreatic Cancer. Available online: https://www.pancan.org/ (accessed on 3 October 2022).
17. Cancer.Net: Hereditary Pancretic Cancer. Available online: https://www.cancer.net/ (accessed on 3 October 2022).
18. BioPortal for Cancer Genomics. Available online: https://www.cbioportal.org/ (accessed on 3 October 2022).
19. Waddell, N.; Pajic, M.; Patch, A.-M.; Chang, D.K.; Kassahn, K.S.; Bailey, P.; Johns, A.L.; Miller, D.; Nones, K.; Quek, K.; et al. Whole Genomes Redefine the Mutational Landscape of Pancreatic Cancer. *Nature* **2015**, *518*, 495–501. https://doi.org/10.1038/nature14169.
20. Pishvaian, M.J.; Bender, R.J.; Halverson, D.; Rahib, L.; Hendifar, A.E.; Mikhail, S.; Chung, V.; Picozzi, V.J.; Sohal, D.; Blais, E.M.; et al. Molecular Profiling of Patients with Pancreatic Cancer: Initial Results from the Know Your Tumor Initiative. *Clin. Cancer Res.* **2018**, *24*, 5018–5027. https://doi.org/10.1158/1078-0432.CCR-18-0531.
21. Dreyer, S.B.; Chang, D.K.; Bailey, P.; Biankin, A. v. Pancreatic Cancer Genomes: Implications for Clinical Management and Therapeutic Development. *Clin. Cancer Res.* **2017**, *23*, 1638–1646. https://doi.org/10.1158/1078-0432.CCR-16-2411.
22. Wattenberg, M.M.; Reiss, K.A. Determinants of Homologous Recombination Deficiency in Pancreatic Cancer. *Cancers* **2021**, *13*, 4716. https://doi.org/10.3390/cancers13184716.
23. Negrini, S.; Gorgoulis, V.G.; Halazonetis, T.D. Genomic Instability—An Evolving Hallmark of Cancer. *Nat. Rev. Mol. Cell Biol.* **2010**, *11*, 220–228. https://doi.org/10.1038/nrm2858.
24. Wang, S.; Zheng, Y.; Yang, F.; Zhu, L.; Zhu, X.-Q.; Wang, Z.-F.; Wu, X.-L.; Zhou, C.-H.; Yan, J.-Y.; Hu, B.-Y.; et al. The Molecular Biology of Pancreatic Adenocarcinoma: Translational Challenges and Clinical Perspectives. *Signal Transduct Target Ther.* **2021**, *6*, 249. https://doi.org/10.1038/s41392-021-00659-4.
25. Felsenstein, M.; Hruban, R.H.; Wood, L.D. New Developments in the Molecular Mechanisms of Pancreatic Tumorigenesis. *Adv. Anat. Pathol.* **2018**, *25*, 131–142. https://doi.org/10.1097/PAP.0000000000000172.
26. Murphy, S.J.; Hart, S.N.; Halling, G.C.; Johnson, S.H.; Smadbeck, J.B.; Drucker, T.; Lima, J.F.; Rohakhtar, F.R.; Harris, F.R.; Kosari, F.; et al. Integrated Genomic Analysis of Pancreatic Ductal Adenocarcinomas Reveals Genomic Rearrangement Events as Significant Drivers of Disease. *Cancer Res.* **2016**, *76*, 749–761. https://doi.org/10.1158/0008-5472.CAN-15-2198.
27. Strzyz, P. Nuclear MRNA Retention Buffers Expression Noise. *Nat. Rev. Mol. Cell Biol.* **2016**, *17*, 67–67. https://doi.org/10.1038/nrm.2016.4.
28. Zhang, X.; Mao, T.; Zhang, B.; Xu, H.; Cui, J.; Jiao, F.; Chen, D.; Wang, Y.; Hu, J.; Xia, Q.; et al. Characterization of the Genomic Landscape in Large-Scale Chinese Patients with Pancreatic Cancer. *EBioMedicine* **2022**, *77*, 103897. https://doi.org/10.1016/j.ebiom.2022.103897.
29. Golan, T.; Hammel, P.; Reni, M.; van Cutsem, E.; Macarulla, T.; Hall, M.J.; Park, J.-O.; Hochhauser, D.; Arnold, D.; Oh, D.-Y.; et al. Maintenance Olaparib for Germline BRCA-Mutated Metastatic Pancreatic Cancer. *N. Engl. J. Med.* **2019**, *381*, 317–327. https://doi.org/10.1056/NEJMoa1903387.
30. O'Reilly, E.M.; Lee, J.W.; Zalupski, M.; Capanu, M.; Park, J.; Golan, T.; Tahover, E.; Lowery, M.A.; Chou, J.F.; Sahai, V.; et al. Randomized, Multicenter, Phase II Trial of Gemcitabine and Cisplatin With or Without Veliparib in Patients With Pancreas Adenocarcinoma and a Germline BRCA/PALB2 Mutation. *J. Clin. Oncol.* **2020**, *38*, 1378–1388. https://doi.org/10.1200/JCO.19.02931.
31. Ogura, T.; Yamao, K.; Hara, K.; Mizuno, N.; Hijioka, S.; Imaoka, H.; Sawaki, A.; Niwa, Y.; Tajika, M.; Kondo, S.; et al. Prognostic Value of K-Ras Mutation Status and Subtypes in Endoscopic Ultrasound-Guided Fine-Needle Aspiration Specimens from Patients with Unresectable Pancreatic Cancer. *J. Gastroenterol.* **2013**, *48*, 640–646. https://doi.org/10.1007/s00535-012-0664-2.
32. Kenney, C.; Kunst, T.; Webb, S.; Christina, D.; Arrowood, C.; Steinberg, S.M.; Mettu, N.B.; Kim, E.J.; Rudloff, U. Phase II Study of Selumetinib, an Orally Active Inhibitor of MEK1 and MEK2 Kinases, in KRASG12R-Mutant Pancreatic Ductal Adenocarcinoma. *Invest. New Drugs* **2021**, *39*, 821–828. https://doi.org/10.1007/s10637-020-01044-8.
33. Furniss, C.S.; Yurgelun, M.B.; Ukaegbu, C.; Constantinou, P.E.; Lafferty, C.C.; Talcove-Berko, E.R.; Schwartz, A.N.; Stopfer, J.E.; Underhill-Blazey, M.; Kenner, B.; et al. Novel Models of Genetic Education and Testing for Pancreatic Cancer Interception: Preliminary Results from the GENERATE Study. *Cancer Prev. Res.* **2021**, *14*, 1021–1032. https://doi.org/10.1158/1940-6207.CAPR-20-0642.
34. Bazzichetto, C.; Luchini, C.; Conciatori, F.; Vaccaro, V.; di Cello, I.; Mattiolo, P.; Falcone, I.; Ferretti, G.; Scarpa, A.; Cognetti, F.; et al. Morphologic and Molecular Landscape of Pancreatic Cancer Variants as the Basis of New Therapeutic Strategies for Precision Oncology. *Int. J. Mol. Sci.* **2020**, *21*, 8841. https://doi.org/10.3390/ijms21228841.

Pobrano z / Downloaded from Repository of Wrocaw University of Environmental and Life Sciences 2025-10-15 21:43

35. Nanda, N.; Roberts, N.J. ATM Serine/Threonine Kinase and Its Role in Pancreatic Risk. *Genes* **2020**, *11*, 108. https://doi.org/10.3390/genes11010108.
36. Klatte, D.C.F.; Wallace, M.B.; Löhr, M.; Bruno, M.J.; van Leerdam, M.E. Hereditary Pancreatic Cancer. *Best Pract. Res. Clin. Gastroenterol.* **2022**, *58–59*, 101783. https://doi.org/10.1016/j.bpg.2021.101783.
37. Tamura, K.; Yu, J.; Hata, T.; Suenaga, M.; Shindo, K.; Abe, T.; MacGregor-Das, A.; Borges, M.; Wolfgang, C.L.; Weiss, M.J.; et al. Mutations in the Pancreatic Secretory Enzymes CPA1 and CPB1 Are Associated with Pancreatic Cancer. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, 4767–4772. https://doi.org/10.1073/pnas.1720588115.
38. Makohon-Moore, A.; Brosnan, J.A.; Iacobuzio-Donahue, C.A. Pancreatic Cancer Genomics: Insights and Opportunities for Clinical Translation. *Genome Med.* **2013**, *5*, 26. https://doi.org/10.1186/gm430.
39. Merz, V.; Mangiameli, D.; Zecchetto, C.; Quinzii, A.; Pietrobono, S.; Messina, C.; Casalino, S.; Gaule, M.; Pesoni, C.; Vitale, P.; et al. Predictive Biomarkers for a Personalized Approach in Resectable Pancreatic Cancer. *Front Surg.* **2022**, *9*, 866173. https://doi.org/10.3389/fsurg.2022.866173.
40. Springer, S.; Wang, Y.; Dal Molin, M.; Masica, D.L.; Jiao, Y.; Kinde, I.; Blackford, A.; Raman, S.P.; Wolfgang, C.L.; Tomita, T.; et al. A Combination of Molecular Markers and Clinical Features Improve the Classification of Pancreatic Cysts. *Gastroenterology* **2015**, *149*, 1501–1510. https://doi.org/10.1053/j.gastro.2015.07.041.
41. Wu, J.; Matthaei, H.; Maitra, A.; Dal Molin, M.; Wood, L.D.; Eshleman, J.R.; Goggins, M.; Canto, M.I.; Schulick, R.D.; Edil, B.H.; et al. Recurrent GNAS Mutations Define an Unexpected Pathway for Pancreatic Cyst Development. *Sci. Transl. Med.* **2011**, *3*, 92ra66. https://doi.org/10.1126/scitranslmed.3002543.
42. Yu, J.; Sadakari, Y.; Shindo, K.; Suenaga, M.; Brant, A.; Almario, J.A.N.; Borges, M.; Barkley, T.; Fesharakizadeh, S.; Ford, M.; et al. Digital Next-Generation Sequencing Identifies Low-Abundance Mutations in Pancreatic Juice Samples Collected from the Duodenum of Patients with Pancreatic Cancer and Intraductal Papillary Mucinous Neoplasms. *Gut* **2017**, *66*, 1677–1687. https://doi.org/10.1136/gutjnl-2015-311166.
43. Bettegowda, C.; Sausen, M.; Leary, R.J.; Kinde, I.; Wang, Y.; Agrawal, N.; Bartlett, B.R.; Wang, H.; Luber, B.; Alani, R.M.; et al. Detection of Circulating Tumor DNA in Early- and Late-Stage Human Malignancies. *Sci. Transl. Med.* **2014**, *6*, 224ra24. https://doi.org/10.1126/scitranslmed.3007094.
44. Tie, J.; Wang, Y.; Tomasetti, C.; Li, L.; Springer, S.; Kinde, I.; Silliman, N.; Tacey, M.; Wong, H.-L.; Christie, M.; et al. Circulating Tumor DNA Analysis Detects Minimal Residual Disease and Predicts Recurrence in Patients with Stage II Colon Cancer. *Sci. Transl. Med.* **2016**, *8*, 346ra92. https://doi.org/10.1126/scitranslmed.aaf6219.
45. Melisi, D.; Cavaliere, A.; Gobbo, S.; Fasoli, G.; Allegrini, V.; Simionato, F.; Gaule, M.; Casalino, S.; Pesoni, C.; Zecchetto, C.; et al. Role of Next-Generation Genomic Sequencing in Targeted Agents Repositioning for Pancreaticoduodenal Cancer Patients. *Pancreatology* **2021**, *21*, 1038–1047. https://doi.org/10.1016/j.pan.2021.04.004.
46. Li, Y.; al Hallak, M.N.; Philip, P.A.; Azmi, A.S.; Mohammad, R.M. Non-Coding RNAs in Pancreatic Cancer Diagnostics and Therapy: Focus on LncRNAs, CircRNAs, and PiRNAs. *Cancers* **2021**, *13*, 4161. https://doi.org/10.3390/cancers13164161.
47. Feigin, M.E.; Garvin, T.; Bailey, P.; Waddell, N.; Chang, D.K.; Kelley, D.R.; Shuai, S.; Gallinger, S.; McPherson, J.D.; Grimmond, S.M.; et al. Recurrent Noncoding Regulatory Mutations in Pancreatic Ductal Adenocarcinoma. *Nat. Genet.* **2017**, *49*, 825–833. https://doi.org/10.1038/ng.3861.
48. Li, J.; Hou, S.; Ye, Z.; Wang, W.; Hu, X.; Hang, Q. Long Non-Coding RNAs in Pancreatic Cancer: Biologic Functions, Mechanisms, and Clinical Significance. *Cancers* **2022**, *14*, 2115. https://doi.org/10.3390/cancers14092115.
49. Campbell, P.J.; Yachida, S.; Mudie, L.J.; Stephens, P.J.; Pleasance, E.D.; Stebbings, L.A.; Morsberger, L.A.; Latimer, C.; McLaren, S.; Lin, M.-L.; et al. The Patterns and Dynamics of Genomic Instability in Metastatic Pancreatic Cancer. *Nature* **2010**, *467*, 1109–1113. https://doi.org/10.1038/nature09460.
50. Yang, Z.-Q.; Liu, Y.-J.; Zhou, X.-L. An Integrated Microarray Analysis Reveals Significant Diagnostic and Prognostic Biomarkers in Pancreatic Cancer. *Med. Sci. Monit.* **2020**, *26*, e921769. https://doi.org/10.12659/MSM.921769.
51. Campagna, D.; Cope, L.; Lakkur, S.S.; Henderson, C.; Laheru, D.; Iacobuzio-Donahue, C.A. Gene Expression Profiles Associated with Advanced Pancreatic Cancer. *Int. J. Clin. Exp. Pathol.* **2008**, *1*, 32–43.
52. Niedergethmann, M.; Alves, F.; Neff, J.K.; Heidrich, B.; Aramin, N.; Li, L.; Pilarsky, C.; Grützmann, R.; Allgayer, H.; Post, S.; et al. Gene Expression Profiling of Liver Metastases and Tumour Invasion in Pancreatic Cancer Using an Orthotopic SCID Mouse Model. *Br. J. Cancer* **2007**, *97*, 1432–1440. https://doi.org/10.1038/sj.bjc.6604031.
53. Yachida, S.; Jones, S.; Bozic, I.; Antal, T.; Leary, R.; Fu, B.; Kamiyama, M.; Hruban, R.H.; Eshleman, J.R.; Nowak, M.A.; et al. Distant Metastasis Occurs Late during the Genetic Evolution of Pancreatic Cancer. *Nature* **2010**, *467*, 1114–1117. https://doi.org/10.1038/nature09515.
54. Maddipati, R.; Stanger, B.Z. Pancreatic Cancer Metastases Harbor Evidence of Polyclonality. *Cancer Discov.* **2015**, *5*, 1086–1097. https://doi.org/10.1158/2159-8290.CD-15-0120.
55. Ro, C.; Chai, W.; Yu, V.E.; Yu, R. Pancreatic Neuroendocrine Tumors: Biology, Diagnosis,and Treatment. *Chin. J. Cancer* **2013**, *32*, 312–324. https://doi.org/10.5732/cjc.012.10295.
56. Petersen, G.M. Familial Pancreatic Cancer. *Semin. Oncol.* **2016**, *43*, 548–553. https://doi.org/10.1053/j.seminoncol.2016.09.002.
57. Hiripi, E.; Lorenzo Bermejo, J.; Li, X.; Sundquist, J.; Hemminki, K. Familial Association of Pancreatic Cancer with Other Malignancies in Swedish Families. *Br. J. Cancer* **2009**, *101*, 1792–1797. https://doi.org/10.1038/sj.bjc.6605363.

Pobrano z / Downloaded from Repository of Wrocaw University of Environmental and Life Sciences 2025-10-15 21:43

58. Cote, M.L.; Schenk, M.; Schwartz, A.G.; Vigneau, F.D.; Kinnard, M.; Greenson, J.K.; Fryzek, J.P.; Ying, G.S.; Garabrant, D.H. Risk of Other Cancers in Individuals with a Family History of Pancreas Cancer. *J. Gastrointest. Cancer* **2007**, *38*, 119–126. https://doi.org/10.1007/s12029-008-9022-2.
59. Klein, A.P. Genetic Susceptibility to Pancreatic Cancer. *Mol. Carcinog.* **2012**, *51*, 14–24. https://doi.org/10.1002/mc.20855.
60. Breast Cancer Linkage Consortium. Cancer Risks in BRCA2 Mutation Carriers. *J. Natl. Cancer Inst.* **1999**, *91*, 1310–1316. https://doi.org/10.1093/jnci/91.15.1310.
61. Giardiello, F.M.; Brensinger, J.D.; Tersmette, A.C.; Goodman, S.N.; Petersen, G.M.; Booker, S. v; Cruz-Correa, M.; Offerhaus, J.A. Very High Risk of Cancer in Familial Peutz-Jeghers Syndrome. *Gastroenterology* **2000**, *119*, 1447–1453. https://doi.org/10.1053/gast.2000.20228.
62. Yang, X.; Leslie, G.; Doroszuk, A.; Schneider, S.; Allen, J.; Decker, B.; Dunning, A.M.; Redman, J.; Scarth, J.; Plaskocinska, I.; et al. Cancer Risks Associated with Germline PALB2 Pathogenic: An International Study of 524 Families. *J. Clin. Oncol.* **2020**, *38*, 674–685. https://doi.org/10.1200/JCO.19.01907.
63. Rustgi, A.K. Familial Pancreatic Cancer: Genetic Advances. *Genes Dev.* **2014**, *28*, 1–7. https://doi.org/10.1101/gad.228452.113.
64. Coco, D.; Leanza, S. Von Hippel-Lindau Syndrome: Medical Syndrome or Surgical Syndrome? A Surgical Perspective. *J. Kidney Cancer VHL* **2021**, *9*, 27–32. https://doi.org/10.15586/jkcvhl.v9i1.206.
65. Ruijs, M.W.G.; Verhoef, S.; Rookus, M.A.; Pruntel, R.; van der Hout, A.H.; Hogervorst, F.B.L.; Kluijt, I.; Sijmons, R.H.; Aalfs Cora, M.; Wagner, A.; et al. TP53 Germline Mutation Testing in 180 Families Suspected Of-Fraumeni Syndrome: Mutation Detection Rate and Relative of Cancers in Different Familial Phenotypes. *J. Med. Genet.* **2010**, *47*, 421–428. https://doi.org/10.1136/jmg.2009.073429.
66. Tischkowitz, M.; Xia, B. PALB2/FANCN: Recombining Cancer and Fanconi Anemia. *Cancer Res.* **2010**, *70*, 7353–7359. https://doi.org/10.1158/0008-5472.CAN-10-1012.
67. Lowenfels, A.B.; Maisonneuve, P.; DiMagno, E.P.; Elitsur, Y.; Gates, L.K.; Perrault, J.; Whitcomb, D.C. Hereditary Pancreatitis and the Risk of Pancreatic Cancer. International Hereditary Pancreatitis Study Group. *J. Natl. Cancer Inst.* **1997**, *89*, 442–446. https://doi.org/10.1093/jnci/89.6.442.
68. Kastrinos, F.; Mukherjee, B.; Tayob, N.; Wang, F.; Sparr, J.; Raymond, V.M.; Bandipalliam, P.; Stoffel, E.M.; Gruber Stephen, B.; Syngal, S. Risk of Pancreatic Cancer in Families with Lynch Syndrome. *JAMA* **2009**, *302*, 1790–1795. https://doi.org/10.1001/jama.2009.1529.
69. Giardiello, F.M.; Offerhaus, G.J.; Lee, D.H.; Krush, A.J.; Tersmette, A.C.; Booker, S. v; Kelley, N.C.; Hamilton, S.R. Increased Risk of Thyroid and Pancreatic Carcinoma in Familial Polyposis. *Gut* **1993**, *34*, 1394–1396. https://doi.org/10.1136/gut.34.10.1394.
70. Gauci, J.; Azzopardi, N.; Babic, D.; Cortis, K.; Axisa, B. Neurofibromatosis Type 1. *Pancreas* **2022**, *51*, 559–562. https://doi.org/10.1097/MPA.0000000000002068.
71. Jensen, R.T.; Berna, M.J.; Bingham, D.B.; Norton, J.A. Inherited Pancreatic Endocrine Tumor Syndromes: Advances in Molecular Pathogenesis, Diagnosis, Management, and Controversies. *Cancer* **2008**, *113*, 1807–1843. https://doi.org/10.1002/cncr.23648.
72. Hu, C.; Hart, S.N.; Polley, E.C.; Gnanaolivu, R.; Shimelis, H.; Lee, K.Y.; Lilyquist, J.; Na, J.; Moore, R.; Antwi, S.O.; et al. Association Between Inherited Germline Mutations in Cancer Predisposition Genes and Risk of Pancreatic Cancer. *JAMA* **2018**, *319*, 2401. https://doi.org/10.1001/jama.2018.6228.
73. Chaffee, K.G.; Oberg, A.L.; McWilliams, R.R.; Majithia, N.; Allen, B.A.; Kidd, J.; Singh, N.; Hartman, A.-R.; Wenstrup, R.J.; Petersen, G.M. Prevalence of Germ-Line Mutations in Cancer Genes among Pancreatic Cancer Patients with a Positive Family History. *Genet. Med.* **2018**, *20*, 119–127. https://doi.org/10.1038/gim.2017.85.
74. Zhen, D.B.; Rabe, K.G.; Gallinger, S.; Syngal, S.; Schwartz, A.G.; Goggins, M.G.; Hruban, R.H.; Cote, M.L.; McWilliams, R.R.; Roberts, N.J.; et al. BRCA1, BRCA2, PALB2, and CDKN2A Mutations in Familial Pancreatic Cancer: A PACGENE Study. *Genet. Med.* **2015**, *17*, 569–577. https://doi.org/10.1038/gim.2014.153.
75. Iqbal, J.; Ragone, A.; Lubinski, J.; Lynch, H.T.; Moller, P.; Ghadirian, P.; Foulkes, W.D.; Armel, S.; Eisen, A.; Neuhausen, S.L.; et al. The Incidence of Pancreatic Cancer in BRCA1 and BRCA2 Carriers. *Br. J. Cancer* **2012**, *107*, 2005–2009. https://doi.org/10.1038/bjc.2012.483.
76. Wong, W.; Raufi, A.G.; Safyan, R.A.; Bates, S.E.; Manji, G.A. BRCA Mutations in Pancreas Cancer: Spectrum, Current, Challenges and Future Prospects. *Cancer Manag. Res.* **2020**, *12*, 2731–2742. https://doi.org/10.2147/CMAR.S211151.
77. Thompson, D.; Easton, D.F.; The Breast Cancer Linkage Consortium. Cancer Incidence in BRCA1 Mutation Carriers. *J. Natl. Cancer Inst.* **2002**, *94*, 1358–1365. https://doi.org/10.1093/jnci/94.18.1358.
78. van Lier, M.G.F.; Wagner, A.; Mathus-Vliegen, E.M.H.; Kuipers, E.J.; Steyerberg, E.W.; van Leerdam, M.E. High Cancer Risk in Peutz–Jeghers Syndrome: A Systematic Review and Surveillance Recommendations. *Am. J. Gastroenterol.* **2010**, *105*, 1258–1264. https://doi.org/10.1038/ajg.2009.725.
79. Korsse, S.E.; Harinck, F.; van Lier, M.G.F.; Biermann, K.; Offerhaus, G.J.A.; Krak, N.; Looman, C.W.N.; van Veelen, W.; Kuipers, E.J.; Wagner, A.; et al. Pancreatic Cancer Risk in Peutz-Jeghers Syndrome Patients: A Large Cohort Study and Implications for Surveillance. *J. Med. Genet.* **2013**, *50*, 59–64. https://doi.org/10.1136/jmedgenet-2012-101277.
80. Lynch, H.T.; Fusaro, R.M.; Lynch, J.F.; Brand, R. Pancreatic Cancer and the FAMMM Syndrome. *Fam. Cancer* **2008**, *7*, 103–112. https://doi.org/10.1007/s10689-007-9166-4.

Pobrano z / Downloaded from Repository of Wrocaw University of Environmental and Life Sciences 2025-10-15 21:43

81. Gu, K.; Velde, R. van de; Pitz, M.; Silver, S. Recurrent Melanoma Development in a Caucasian Female with CDKN2A+ Mutation and FAMMM Syndrome: A Case Report. SAGE Open Med. *Case Rep.* **2020**, *8*, 2050313X20936034. https://doi.org/10.1177/2050313X20936034.
82. Zwolak, A.; Świrska, J.; Tywanek, E.; Dudzińska, M.; Tarach, J.S.; Matyjaszek-Matuszek, B. Pancreatic Neuroendocrine Tumours in Patients with Von-Lindau Disease. *Endokrynol. Pol.* **2020**, *71*, 256–259. https://doi.org/10.5603/EP.a2020.0027.
83. Mezquita, L.; Jové, M.; Nadal, E.; Kfoury, M.; Morán, T.; Ricordel, C.; Dhooge, M.; Tlemsani, C.; Léna, H.; Teulé, A.; et al. High Prevalence of Somatic Oncogenic Driver Alterations in Patients With NSCLC and Li-Fraumeni Syndrome. *J. Thorac. Oncol.* **2020**, *15*, 1232–1239. https://doi.org/10.1016/j.jtho.2020.03.005.
84. Roberts, N.J.; Norris, A.L.; Petersen, G.M.; Bondy, M.L.; Brand, R.; Gallinger, S.; Kurtz, R.C.; Olson, S.H.; Rustgi, A.K.; Schwartz, A.G.; et al. Whole Genome Sequencing Defines the Genetic Heterogeneity of Familial Pancreatic Cancer. *Cancer Discov.* **2016**, *6*, 166–175. https://doi.org/10.1158/2159-8290.CD-15-0402.
85. Howes, N.; Lerch, M.M.; Greenhalf, W.; Stocken, D.D.; Ellis, I.; Simon, P.; Truninger, K.; Ammann, R.; Cavallini, G.; Charnley, R.M.; et al. Clinical and Genetic Characteristics of Hereditary Pancreatitis in Europe. *Clin. Gastroenterol. Hepatol.* **2004**, *2*, 252–261. https://doi.org/10.1016/S1542-3565(04)00013-8.
86. Hes, F.J.; Nielsen, M.; Bik, E.C.; Konvalinka, D.; Wijnen, J.T.; Bakker, E.; Vasen, H.F.A.; Breuning, M.H.; Tops, C.M.J. Somatic APC Mosaicism: An Underestimated Cause of Polyposis. *Gut* **2008**, *57*, 71–76. https://doi.org/10.1136/gut.2006.117796.
87. Burt, R.W. Colon Cancer Screening. *Gastroenterology* **2000**, *119*, 837–853. https://doi.org/10.1053/gast.2000.16508.
88. Møller, P.; Seppälä, T.T.; Bernstein, I.; Holinski-Feder, E.; Sala, P.; Gareth Evans, D.; Lindblom, A.; Macrae, F.; Blanco, I.; Sijmons, R.H.; et al. Cancer Risk and Survival in Path_MMR Carriers by Gene and Gender up to 75 Years of Age: A Report from the Prospective Lynch Syndrome Database. *Gut* **2018**, *67*, 1306–1316. https://doi.org/10.1136/gutjnl-2017-314057.
89. Haraldsdottir, S.; Rafnar, T.; Frankel, W.L.; Einarsdottir, S.; Sigurdsson, A.; Hampel, H.; Snaebjornsson, P.; Masson, G.; Weng, D.; Arngrimsson, R.; et al. Comprehensive Population-Wide Analysis of Lynch Syndrome in Iceland Reveals Founder Mutations in MSH6 and PMS2. *Nat. Commun.* **2017**, *8*, 14755. https://doi.org/10.1038/ncomms14755.
90. Venkitaraman, A.R. Cancer Susceptibility and the Functions of BRCA1 and BRCA2. *Cell* **2002**, *108*, 171–182. https://doi.org/10.1016/S0092-8674(02)00615-3.
91. Brose, M.S.; Rebbeck, T.R.; Calzone, K.A.; Stopfer, J.E.; Nathanson, K.L.; Weber, B.L. Cancer Risk Estimates for BRCA1 Mutation Carriers Identified in a Risk Evaluation Program. *J. Natl. Cancer Inst.* **2002**, *94*, 1365–1372. https://doi.org/10.1093/jnci/94.18.1365.
92. Moran, A.; O'Hara, C.; Khan, S.; Shack, L.; Woodward, E.; Maher, E.R.; Lalloo, F.; Evans, D.G.R. Risk of Cancer Other than Breast or Ovarian in Individuals with BRCA1 and BRCA2 Mutations. *Fam. Cancer* **2012**, *11*, 235–242. https://doi.org/10.1007/s10689-011-9506-2.
93. Mersch, J.; Jackson, M.A.; Park, M.; Nebgen, D.; Peterson, S.K.; Singletary, C.; Arun, B.K.; Litton, J.K. Cancers Associated with BRCA1 and BRCA2 Mutations Other than Breast and Ovarian. *Cancer* **2015**, *121*, 269–275. https://doi.org/10.1002/cncr.29041.
94. Ferrone, C.R.; Levine, D.A.; Tang, L.H.; Allen, P.J.; Jarnagin, W.; Brennan, M.F.; Offit, K.; Robson, M.E. BRCA Germline Mutations in Jewish Patients With Pancreatic Adenocarcinoma. *J. Clin. Oncol.* **2009**, *27*, 433–438. https://doi.org/10.1200/JCO.2008.18.5546.
95. Marmolejo, D.H.; Wong, M.Y.Z.; Bajalica-Lagercrantz, S.; Tischkowitz, M.; Balmaña, J.; Patócs, A.B.; Chappuis, P.; Colas, C.; Genuardi, M.; Haanpää, M.; et al. Overview of Hereditary Breast and Ovarian Cancer (HBOC) Guidelines across Europe. *Eur. J. Med. Genet.* **2021**, *64*, 104350. https://doi.org/10.1016/j.ejmg.2021.104350.
96. Bergman, W.; Watson, P.; de Jong, J.; Lynch, H.; Fusaro, R. Systemic Cancer and the FAMMM Syndrome. *Br. J. Cancer* **1990**, *61*, 932–936. https://doi.org/10.1038/bjc.1990.209.
97. Hansson, J. Familial Cutaneous Melanoma. In *Diseases of DNA Repair*; Springer: New York, NY, USA, 2010; pp. 134–145.
98. Lynch, H.T.; Shaw, T.G. Familial Atypical Multiple Mole Melanoma (FAMMM) Syndrome: History, Genetics, and Heterogeneity. *Fam. Cancer* **2016**, *15*, 487–491. https://doi.org/10.1007/s10689-016-9888-2.
99. Das, K.K.; Early, D. Pancreatic Cancer Screening. *Curr. Treat. Options Gastroenterol.* **2017**, *15*, 562–575. https://doi.org/10.1007/s11938-017-0149-8.
100. Lynch, H.T.; Brand, R.E.; Hogg, D.; Deters, C.A.; Fusaro, R.M.; Lynch, J.F.; Liu, L.; Knezetic, J.; Lassam, N.J.; Goggins, M.; et al. Phenotypic Variation in Eight ExtendedCDKN2A Germline Mutation Familial Atypical Multiple Mole Melanoma-Pancreatic Carcinoma-Prone Families. *Cancer* **2002**, *94*, 84–96. https://doi.org/10.1002/cncr.10159.
101. Goldstein, A.M.; Fraser, M.C.; Struewing, J.P.; Hussussian, C.J.; Ranade, K.; Zametkin, D.P.; Fontaine, L.S.; Organic, S.M.; Dracopoli, N.C.; Clark, W.H.; et al. Increased Risk of Pancreatic Cancer in Melanoma-Prone Kindreds with P16INK4 Mutations. *N. Engl. J. Med.* **1995**, *333*, 970–975. https://doi.org/10.1056/NEJM199510123331504.
102. Abe, K.; Kitago, M.; Kitagawa, Y.; Hirasawa, A. Hereditary Pancreatic Cancer. *Int. J. Clin. Oncol.* **2021**, *26*, 1784–1792. https://doi.org/10.1007/s10147-021-02015-6.
103. Bujanda, L.; Herreros-Villanueva, M. Pancreatic Cancer in Lynch Syndrome Patients. *J. Cancer* **2017**, *8*, 3667–3674. https://doi.org/10.7150/jca.20750.
104. Seppälä, T.T.; Latchford, A.; Negoi, I.; Sampaio Soares, A.; Jimenez-Rodriguez, R.; Sánchez-Guillén, L.; Evans, D.G.; Ryan, N.; Crosbie, E.J.; Dominguez-Valentin, M.; et al. European Guidelines from the EHTG and ESCP for Lynch Syndrome: An Updated

Pobrano z / Downloaded from Repository of Wrocaw University of Environmental and Life Sciences 2025-10-15 21:43

Third Edition of the Mallorca Guidelines Based on Gene and Gender. *Br. J. Surg.* **2021**, *108*, 484–498. https://doi.org/10.1002/bjs.11902.

105. Steinke, V.; Engel, C.; Büttner, R.; Schackert, H.K.; Schmiegel, W.H.; Propping, P. Hereditary Nonpolyposis Colorectal Cancer (HNPCC)/Lynch Syndrome. *Dtsch. Arztebl. Int.* **2013**, *110*, 32–38. https://doi.org/10.3238/arztebl.2013.0032.
106. Tacheci, I.; Kopacova, M.; Bures, J. Peutz-Jeghers Syndrome. *Curr. Opin. Gastroenterol.* **2021**, *37*, 245–254. https://doi.org/10.1097/MOG.0000000000000718.
107. Signoretti, M.; Bruno, M.J.; Zerboni, G.; Poley, J.-W.; Delle Fave, G.; Capurso, G. Results of Surveillance in Individuals at High-risk of Pancreatic Cancer: A Systematic Review and Meta-analysis. *United Eur. Gastroenterol. J.* **2018**, *6*, 489–499. https://doi.org/10.1177/2050640617752182.
108. Spoto, C.P.E.; Gullo, I.; Carneiro, F.; Montgomery, E.A.; Brosens, L.A.A. Hereditary Gastrointestinal Carcinomas and Their Precursors: An Algorithm for Genetic Testing. *Semin. Diagn. Pathol.* **2018**, *35*, 170–183. https://doi.org/10.1053/j.semdp.2018.01.004.
109. Benzel, J.; Fendrich, V. Familial Pancreatic Cancer. *Oncol. Res. Treat.* **2018**, *41*, 611–618. https://doi.org/10.1159/000493473.
110. Guha, T.; Malkin, D. Inherited TP53 Mutations and the Li–Fraumeni Syndrome. *Cold Spring Harb. Perspect. Med.* **2017**, *7*, a026187. https://doi.org/10.1101/cshperspect.a026187.
111. Mai, P.L.; Best, A.F.; Peters, J.A.; DeCastro, R.M.; Khincha, P.P.; Loud, J.T.; Bremer, R.C.; Rosenberg, P.S.; Savage, S.A. Risks of First and Subsequent Cancers among TP53 Mutation Carriers in the National Cancer Institute Li-Fraumeni Syndrome Cohort. *Cancer* **2016**, *122*, 3673–3681. https://doi.org/10.1002/cncr.30248.
112. Waller, A.; Findeis, S.; Lee, M. Familial Adenomatous Polyposis. *J. Pediatr. Genet.* **2016**, *5*, 78–83. https://doi.org/10.1055/s-0036-1579760.
113. Petersen, G.M.; Slack, J.; Nakamura, Y. Screening Guidelines and Premorbid Diagnosis of Familial Adenomatous Polyposis Using Linkage. *Gastroenterology* **1991**, *100*, 1658–1664. https://doi.org/10.1016/0016-5085(91)90666-9.
114. Chintalacheruvu, L.M.; Shaw, T.; Buddam, A.; Diab, O.; Kassim, T.; Mukherjee, S.; Lynch, H.T. Major Hereditary Gastrointestinal Cancer Syndromes: A Narrative Review. *J. Gastrointest. Liver Dis.* **2017**, *26*, 157–163. https://doi.org/10.15403/jgld.2014.1121.262.maj.
115. Dinarvand, P.; Davaro, E.P.; Doan, J. v.; Ising, M.E.; Evans, N.R.; Phillips, N.J.; Lai, J.; Guzman, M.A. Familial Adenomatous Polyposis Syndrome: An Update and Review of Extraintestinal Manifestations. *Arch. Pathol. Lab. Med.* **2019**, *143*, 1382–1398. https://doi.org/10.5858/arpa.2018-0570-RA.
116. Kanth, P.; Grimmett, J.; Champine, M.; Burt, R.; Samadder, J.N. Hereditary Colorectal Polyposis and Cancer Syndromes: A Primer on Diagnosis and Management. *Am. J. Gastroenterol.* **2017**, *112*, 1509–1525. https://doi.org/10.1038/ajg.2017.212.
117. Horii, A.; Nakatsuru, S.; Miyoshi, Y.; Ichii, S.; Nagase, H.; Ando, H.; Yanagisawa, A.; Tsuchiya, E.; Kato, Y.; Nakamura, Y. Frequent Somatic Mutations of the APC Gene in Human Pancreatic Cancer. *Cancer Res.* **1992**, *52*, 6696–6698.
118. Ding, S.-F.; Habib, N.; Delhanty, J.; Bowles, L.; Greco, L.; Wood, C.; Williamson, R.; Dooley, J. Loss of Heterozygosity on Chromosomes 1 and 11 in Carcinoma of the Pancreas. *Br. J. Cancer* **1992**, *65*, 809–812. https://doi.org/10.1038/bjc.1992.173.
119. Neuman, W.L.; Wasylyshyn, M.L.; Jacoby, R.; Erroi, F.; Angriman, I.; Montag, A.; Brasitus, T.; Michelassi, F.; Westbrook, C.A. Evidence for a Common Molecular Pathogenesis in Colorectal, Gastric, and Pancreatic Cancer. *Genes Chromosomes Cancer* **1991**, *3*, 468–473. https://doi.org/10.1002/gcc.2870030609.
120. Roberts, N.J.; Jiao, Y.; Yu, J.; Kopelovich, L.; Petersen, G.M.; Bondy, M.L.; Gallinger, S.; Schwartz, A.G.; Syngal, S.; Cote, M.L.; et al. ATM Mutations in Patients with Hereditary Pancreatic Cancer. *Cancer Discov.* **2012**, *2*, 41–46. https://doi.org/10.1158/2159-8290.CD-11-0194.
121. Savitsky, K.; Bar-Shira, A.; Gilad, S.; Rotman, G.; Ziv, Y.; Vanagaite, L.; Tagle, D.A.; Smith, S.; Uziel, T.; Sfez, S.; et al. A Single Ataxia Telangiectasia Gene with a Product Similar to PI-3 Kinase. *Science* **1995**, *268*, 1749–1753. https://doi.org/10.1126/science.7792600.
122. Jones, S.; Zhang, X.; Parsons, D.W.; Lin, J.C.-H.; Leary, R.J.; Angenendt, P.; Mankoo, P.; Carter, H.; Kamiyama, H.; Jimeno, A.; et al. Core Signaling Pathways in Human Pancreatic Cancers Revealed by Global Genomic Analyses. *Science* **2008**, *321*, 1801–1806. https://doi.org/10.1126/science.1164368.
123. Russell, R.; Perkhofer, L.; Liebau, S.; Lin, Q.; Lechel, A.; Feld, F.M.; Hessmann, E.; Gaedcke, J.; Güthle, M.; Zenke, M.; et al. Loss of ATM Accelerates Pancreatic Cancer Formation and Epithelial–Mesenchymal Transition. *Nat. Commun.* **2015**, *6*, 7677. https://doi.org/10.1038/ncomms8677.
124. Murphy, S.J.; Hart, S.N.; Lima, J.F.; Kipp, B.R.; Klebig, M.; Winters, J.L.; Szabo, C.; Zhang, L.; Eckloff, B.W.; Petersen, G.M.; et al. Genetic Alterations Associated With Progression From Pancreatic Intraepithelial Neoplasia to Invasive Pancreatic Tumor. *Gastroenterology* **2013**, *145*, 1098–1109.e1. https://doi.org/10.1053/j.gastro.2013.07.049.
125. Hsu, F.-C.; Roberts, N.J.; Childs, E.; Porter, N.; Rabe, K.G.; Borgida, A.; Ukaegbu, C.; Goggins, M.G.; Hruban, R.H.; Zogopoulos, G.; et al. Risk of Pancreatic Cancer Among Individuals With Pathogenic Variants in the ATM Gene. *JAMA Oncol.* **2021**, *7*, 1664. https://doi.org/10.1001/jamaoncol.2021.3701.
126. Michl, J.; Zimmer, J.; Tarsounas, M. Interplay between Fanconi Anemia and Homologous Recombination Pathways in Genome Integrity. *EMBO J.* **2016**, *35*, 909–923. https://doi.org/10.15252/embj.201693860.

Pobrano z / Downloaded from Repository of Wrocaw University of Environmental and Life Sciences 2025-10-15 21:43

127. Smith, A.L.; Alirezaie, N.; Connor, A.; Chan-Seng-Yue, M.; Grant, R.; Selander, I.; Bascuñana, C.; Borgida, A.; Hall, A.; Whelan, T.; et al. Candidate DNA Repair Susceptibility Genes Identified by Exome Sequencing in High-Risk Pancreatic Cancer. *Cancer Lett.* **2016**, *370*, 302–312. https://doi.org/10.1016/j.canlet.2015.10.030.
128. Rogers, C.D.; van der Heijden, M.S.; Brune, K.; Yeo, C.J.; Hruban, R.H.; Kern, S.E.; Goggins, M. The Genetics of FANCC and FANCG in Familial Pancreatic Cancer. *Cancer Biol. Ther.* **2004**, *3*, 167–169. https://doi.org/10.4161/cbt.3.2.609.
129. van der Heijden, M.S.; Brody, J.R.; Gallmeier, E.; Cunningham, S.C.; Dezentje, D.A.; Shen, D.; Hruban, R.H.; Kern, S.E. Functional Defects in the Fanconi Anemia Pathway in Pancreatic Cancer Cells. *Am. J. Pathol.* **2004**, *165*, 651–657. https://doi.org/10.1016/S0002-9440(10)63329-9.
130. Mathew, C.G. Fanconi Anaemia Genes and Susceptibility to Cancer. *Oncogene* **2006**, *25*, 5875–5884. https://doi.org/10.1038/sj.onc.1209878.
131. Couch, F.J.; Johnson, M.R.; Rabe, K.; Boardman, L.; McWilliams, R.; de Andrade, M.; Petersen, G. Germ Line Fanconi Anemia Complementation Group C Mutations and Pancreatic Cancer. *Cancer Res.* **2005**, *65*, 383–386.
132. Newey, P.J.; Thakker, R. v. Role of Multiple Endocrine Neoplasia Type 1 Mutational Analysis in Clinical Practice. *Endocr. Pract.* **2011**, *17*, 8–17. https://doi.org/10.4158/EP10379.RA.
133. Al-Salameh, A.; Baudry, C.; Cohen, R. Update on Multiple Endocrine Neoplasia Type 1 and 2. *Presse Med.* **2018**, *47*, 722–731. https://doi.org/10.1016/j.lpm.2018.03.005.
134. Ito, T.; Sasano, H.; Tanaka, M.; Osamura, R.Y.; Sasaki, I.; Kimura, W.; Takano, K.; Obara, T.; Ishibashi, M.; Nakao, K.; et al. Epidemiological Study of Gastroenteropancreatic Neuroendocrine Tumors in Japan. *J. Gastroenterol.* **2010**, *45*, 234–243. https://doi.org/10.1007/s00535-009-0194-8.
135. Thakker, R. v.; Newey, P.J.; Walls, G. v.; Bilezikian, J.; Dralle, H.; Ebeling, P.R.; Melmed, S.; Sakurai, A.; Tonelli, F.; Brandi, M.L. Clinical Practice Guidelines for Multiple Endocrine Neoplasia Type 1 (MEN1). *J. Clin. Endocrinol. Metab.* **2012**, *97*, 2990–3011. https://doi.org/10.1210/jc.2012-1230.
136. Thakker, R. v. Multiple Endocrine Neoplasia Type 1 (MEN1) and Type 4 (MEN4). *Mol. Cell Endocrinol.* **2014**, *386*, 2–15. https://doi.org/10.1016/j.mce.2013.08.002.
137. Kfir, S.K.; Halperin, R.; Percik, R.; Uri, I.; Halpern, N.; Shlomai, G.; Laish, I.; Tirosh, A.; Tirosh, A. Distinct Prognostic Factors in Sporadic and Multiple Endocrine Neoplasia Type 1-Related Pancreatic Neuroendocrine Tumors. *Horm. Metab. Res.* **2021**, *53*, 319–325. https://doi.org/10.1055/a-1464-1276.
138. Tsang, S.H.; Sharma, T. Von Hippel-Lindau Disease. In *Atlas of Inherited Retinal Diseases*; Springer: Cham, Switzerland, 2018; pp. 201–203.
139. Varshney, N.; Kebede, A.A.; Owusu-Dapaah, H.; Lather, J.; Kaushik, M.; Bhullar, J.S. A Review of Von Hippel-Lindau Syndrome. *J. Kidney Cancer VHL* **2017**, *4*, 20–29. https://doi.org/10.15586/jkcvhl.2017.88.
140. Hammel, P.R.; Vilgrain, V.; Terris, B.; Penfornis, A.; Sauvanet, A.; Correas, J.; Chauveau, D.; Balian, A.; Beigelman, C.; O'Toole, D.; et al. Pancreatic Involvement in von Hippel–Lindau Disease. *Gastroenterology* **2000**, *119*, 1087–1095. https://doi.org/10.1053/gast.2000.18143.
141. Reich, M.; Jaegle, S.; Neumann-Haefelin, E.; Klingler, J.; Evers, C.; Daniel, M.; Bucher, F.; Ludwig, F.; Nuessle, S.; Kopp, J.; et al. Genotype–Phenotype Correlation in von Hippel-Lindau Disease. *Acta Ophthalmol.* **2021**, *99*, e1492–e1500. https://doi.org/10.1111/aos.14843.
142. Ferner, R.E.; Huson, S.M.; Thomas, N.; Moss, C.; Willshaw, H.; Evans, D.G.; Upadhyaya, M.; Towers, R.; Gleeson, M.; Steiger, C.; et al. Guidelines for the Diagnosis and Management of Individuals with Neurofibromatosis 1. *J. Med. Genet.* **2006**, *44*, 81–88. https://doi.org/10.1136/jmg.2006.045906.
143. Cawthon, R.M.; Weiss, R.; Xu, G.; Viskochil, D.; Culver, M.; Stevens, J.; Robertson, M.; Dunn, D.; Gesteland, R.; O'Connell, P.; et al. A Major Segment of the Neurofibromatosis Type 1 Gene: CDNA Sequence, Genomic Structure, and Point Mutations. *Cell* **1990**, *62*, 193–201. https://doi.org/10.1016/0092-8674(90)90253-B.
144. Nishi, T.; Kawabata, Y.; Hari, Y.; Imaoka, H.; Ishikawa, N.; Yano, S.; Maruyama, R.; Tajima, Y. A Case of Pancreatic Neuroendocrine Tumor in a Patient with Neurofibromatosis-1. *World J. Surg. Oncol.* **2012**, *10*, 153. https://doi.org/10.1186/1477-7819-10-153.
145. Fujisawa, T.; Osuga, T.; Maeda, M.; Sakamoto, N.; Maeda, T.; Sakaguchi, K.; Onishi, Y.; Toyoda, M.; Maeda, H.; Miyamoto, K.; et al. Malignant Endocrine Tumor of the Pancreas Associated with von Recklinghausen's Disease. *J. Gastroenterol.* **2002**, *37*, 59–67. https://doi.org/10.1007/s535-002-8135-x.
146. Schwartz, R.A.; Fernández, G.; Kotulska, K.; Jóźwiak, S. Tuberous Sclerosis Complex: Advances in Diagnosis, Genetics, and Management. *J. Am. Acad. Dermatol.* **2007**, *57*, 189–202. https://doi.org/10.1016/j.jaad.2007.05.004.
147. Sepp, T.; Yates, J.R.; Green, A.J. Loss of Heterozygosity in Tuberous Sclerosis Hamartomas. *J. Med. Genet.* **1996**, *33*, 962–964. https://doi.org/10.1136/jmg.33.11.962.
148. Cheadle, J.P.; Reeve, M.P.; Sampson, J.R.; Kwiatkowski, D.J. Molecular Genetic Advances in Tuberous Sclerosis. *Hum. Genet.* **2000**, *107*, 97–114. https://doi.org/10.1007/s004390000348.
149. Merritt, J.L.; Davis, D.M.R.; Pittelkow, M.R.; Babovic-Vuksanovic, D. Extensive Acrochordons and Pancreatic Islet-Cell Tumors in Tuberous Sclerosis Associated WithTSC2 Mutations. *Am. J. Med. Genet. A* **2006**, *140A*, 1669–1672. https://doi.org/10.1002/ajmg.a.31351.

Pobrano z / Downloaded from Repository of Wrocaw University of Environmental and Life Sciences 2025-10-15 21:43

150. Francalanci, P.; Diomedi-Camassei, F.; Purificato, C.; Santorelli, F.M.; Giannotti, A.; Dominici, C.; Inserra, A.; Boldrini, R. Malignant Pancreatic Endocrine Tumor in a Child With Tuberous Sclerosis. *Am. J. Surg. Pathol.* **2003**, *27*, 1386–1389. https://doi.org/10.1097/00000478-200310000-00012.
151. Verhoef, S.; van Diemen-Steenvoorde, R.; Akkersdijk, W.L.; Bax, N.M.A.; Ariyurek, Y.; Hermans, C.J.; van Nieuwenhuizen, O.; Nikkels, P.G.J.; Lindhout, D.; Halley, D.J.J.; et al. Malignant Pancreatic Tumour within the Spectrum of Tuberous Sclerosis Complex in Childhood. *Eur. J. Pediatr.* **1999**, *158*, 284–287. https://doi.org/10.1007/s004310051073.
152. Wood, L.D.; Canto, M.I.; Jaffee, E.M.; Simeone, D.M. Pancreatic Cancer: Pathogenesis, Screening, Diagnosis, and Treatment. *Gastroenterology* **2022**, *163*, 386-402.e1. https://doi.org/10.1053/j.gastro.2022.03.056.
153. Zhao, Z.; Liu, W. Pancreatic Cancer: A Review of Risk Factors, Diagnosis, and Treatment. *Technol. Cancer Res. Treat.* **2020**, *19*, 153303382096211. https://doi.org/10.1177/1533033820962117.
154. Chen, H.; Zhuo, Q.; Ye, Z.; Xu, X.; Ji, S. Organoid Model: A New Hope for Pancreatic Cancer Treatment? *Biochim. Biophys. Acta (BBA)-Rev. Cancer* **2021**, *1875*, 188466. https://doi.org/10.1016/j.bbcan.2020.188466.
155. Blomstrand, H.; Batra, A.; Cheung, W.Y.; Elander, N.O. Real-World Evidence on First- and Second-Line Palliative Chemotherapy in Advanced Pancreatic Cancer. *World J. Clin. Oncol.* **2021**, *12*, 787–799. https://doi.org/10.5306/wjco.v12.i9.787.
156. Golan, T.; Kanji, Z.S.; Epelbaum, R.; Devaud, N.; Dagan, E.; Holter, S.; Aderka, D.; Paluch-Shimon, S.; Kaufman, B.; Gershoni-Baruch, R.; et al. Overall Survival and Clinical Characteristics of Pancreatic Cancer in BRCA Mutation Carriers. *Br. J. Cancer* **2014**, *111*, 1132–1138. https://doi.org/10.1038/bjc.2014.418.
157. Tempero, M.A. NCCN Guidelines Updates: Pancreatic Cancer. *J. Natl. Compr. Canc. Netw.* **2019**, *17*, 603–605. https://doi.org/10.6004/jnccn.2019.5007.
158. Mirza, M.R.; Monk, B.J.; Herrstedt, J.; Oza, A.M.; Mahner, S.; Redondo, A.; Fabbro, M.; Ledermann, J.A.; Lorusso, D.; Vergote, I.; et al. Niraparib Maintenance Therapy in Platinum-Sensitive, Recurrent Ovarian Cancer. *N. Engl. J. Med.* **2016**, *375*, 2154–2164. https://doi.org/10.1056/NEJMoa1611310.
159. Mishra, N.K.; Southekal, S.; Guda, C. Survival Analysis of Multi-Omics Data Identifies Potential Prognostic Markers of Pancreatic Ductal Adenocarcinoma. *Front Genet.* **2019**, *10*, 624. https://doi.org/10.3389/fgene.2019.00624.
160. Shi, H.; Xu, H.; Chai, C.; Qin, Z.; Zhou, W. Integrated Bioinformatics Analysis of Potential Biomarkers for Pancreatic Cancer. *J. Clin. Lab. Anal.* **2022**, *36*, e24381. https://doi.org/10.1002/jcla.24381.
161. Sakharkar, M.K.; Dhillon, S.K.; Mazumder, M.; Yang, J. Key Drug-Targeting Genes in Pancreatic Ductal Adenocarcinoma. *Genes Cancer* **2021**, *12*, 12–24. https://doi.org/10.18632/genesandcancer.210.
162. Daoud, A.Z.; Mulholland, E.J.; Cole, G.; McCarthy, H.O. MicroRNAs in Pancreatic Cancer: Biomarkers, Prognostic, and Therapeutic Modulators. *BMC Cancer* **2019**, *19*, 1130. https://doi.org/10.1186/s12885-019-6284-y.
163. Bayer A Study to Test the Effect of the Drug Larotrectinib in Adults and Children With NTRK-Fusion Positive Solid Tumors. Available online: https://clinicaltrials.gov/ct2/show/NCT02576431 (accessed on 10 January 2023)
164. O'Reilly, E.M.; Hechtman, J.F. Tumour Response to TRK Inhibition in a Patient with Pancreatic Adenocarcinoma Harbouring an NTRK Gene Fusion. *Ann. Oncol.* **2019**, *30*, viii36–viii40. https://doi.org/10.1093/annonc/mdz385.
165. Torres, C.; Grippo, P.J. Pancreatic Cancer Subtypes: A Roadmap for Precision Medicine. *Ann. Med.* **2018**, *50*, 277–287. https://doi.org/10.1080/07853890.2018.1453168.
166. Nishimura, M.; Matsukawa, M.; Fujii, Y.; Matsuda, Y.; Arai, T.; Ochiai, Y.; Itoi, T.; Yahagi, N. Effects of EUS-Guided Intratumoral Injection of Oligonucleotide STNM01 on Tumor Growth, Histology, and Overall Survival in Patients with Unresectable Pancreatic Cancer. *Gastrointest. Endosc.* **2018**, *87*, 1126–1131. https://doi.org/10.1016/j.gie.2017.10.030.
167. Yamakawa, K.; Nakano-Narusawa, Y.; Hashimoto, N.; Yokohira, M.; Matsuda, Y. Development and Clinical Trials of Nucleic Acid Medicines for Pancreatic Cancer Treatment. *Int. J. Mol. Sci.* **2019**, *20*, 4224. https://doi.org/10.3390/ijms20174224.
168. Ko, A. Erlotinib in the Treatment of Advanced Pancreatic Cancer. *Biologics* **2008**, *2*, 83–95. https://doi.org/10.2147/BTT.S1832.
169. Hoyer, K.; Hablesreiter, R.; Inoue, Y.; Yoshida, K.; Briest, F.; Christen, F.; Kakiuchi, N.; Yoshizato, T.; Shiozawa, Y.; Shiraishi, Y.; et al. A Genetically Defined Signature of Responsiveness to Erlotinib in Early-Stage Pancreatic Cancer Patients: Results from the CONKO-005 Trial. *EBioMedicine* **2021**, *66*, 103327. https://doi.org/10.1016/j.ebiom.2021.103327.
170. Han, B.; Kim, B.J.; Kim, H.S.; Choi, D.R.; Shim, B.Y.; Lee, K.H.; Kim, J.W.; Kim, J.H.; Song, H.; Kim, J.H.; et al. A Phase II Study of Gemcitabine, Erlotinib and S-1 in Patients with Advanced Pancreatic Cancer. *J. Cancer* **2021**, *12*, 912–917. https://doi.org/10.7150/jca.50514.
171. U. S. Food and Drug Administration Everolimus (Afinitor). Available online: https://www.fda.gov/drugs/resources-information-approved-drugs/everolimus-afinitor (accessed on 10 January 2023).
172. Drugs.com Belzutifan. Available online: https://www.drugs.com/mtm/belzutifan.html (accessed on 10 January 2023).
173. Neuroendocrine Tumor Research Foundation Clinical Trial of Belzutifan in Pancreatic NETs and Pheo/Para. Available online: https://netrf.org/2021/07/27/clinical-trial-of-belzutifan-in-pancreatic-nets-and-pheo-para/ (accessed on 10 January 2023).
174. Pelle, E.; Al-Toubah, T.; Morse, B.; Strosberg, J. Belzutifan in a Patient With VHL-Associated Metastatic Pancreatic Neuroendocrine Tumor. *J. Natl. Compr. Cancer Netw.* **2022**, *20*, 1285–1287. https://doi.org/10.6004/jnccn.2022.7047.
175. Varghese, A.; O'Reilly, E. An Open-Label, Multicenter, Phase 2 Study to Evaluate the Efficacy and Safety of Pembrolizumab Plus Lenvatinib in Combination with Belzutifan in Multiple Solid Tumors. Available online: https://www.mskcc.org/cancer-care/clinical-trials/22-169 (accessed on 10 January 2023).

Pobrano z / Downloaded from Repository of Wrocaw University of Environmental and Life Sciences 2025-10-15 21:43

176. al Baghdadi, T.; Halabi, S.; Garrett-Mayer, E.; Mangat, P.K.; Ahn, E.R.; Sahai, V.; Alvarez, R.H.; Kim, E.S.; Yost, K.J.; Rygiel, A.L.; et al. Palbociclib in Patients With Pancreatic and Biliary Cancer With CDKN2A Alterations: Results From the Targeted Agent and Profiling Utilization Registry Study. *JCO Precis Oncol.* **2019**, *3*, 1–8. https://doi.org/10.1200/PO.19.00124.
177. Huang, H.; Zhang, Y.; Gallegos, V.; Sorrelle, N.; Zaid, M.M.; Toombs, J.; Du, W.; Wright, S.; Hagopian, M.; Wang, Z.; et al. Targeting TGFβR2-mutant Tumors Exposes Vulnerabilities to Stromal TGFβ Blockade in Pancreatic Cancer. *EMBO Mol. Med.* **2019**, *11*, e10515. https://doi.org/10.15252/emmm.201910515.
178. Chen, Y.; Xu, S.; Liu, X.; Jiang, X.; Jiang, J. CircSEC24A Upregulates TGFBR2 Expression to Accelerate Pancreatic Cancer Proliferation and Migration via Sponging to MiR-606. *Cancer Cell Int.* **2021**, *21*, 671. https://doi.org/10.1186/s12935-021-02392-y.
179. Shugang, X.; Hongfa, Y.; Jianpeng, L.; Xu, Z.; Jingqi, F.; Xiangxiang, L.; Wei, L. Prognostic Value of SMAD4 in Pancreatic Cancer: A Meta-Analysis. *Transl. Oncol.* **2016**, *9*, 1–7. https://doi.org/10.1016/j.tranon.2015.11.007.
180. Wang, L.; Liu, Z.; Zhu, R.; Liang, R.; Wang, W.; Li, J.; Zhang, Y.; Guo, C.; Han, X.; Sun, Y. Multi-Omics Landscape and Clinical Significance of a SMAD4-Driven Immune Signature: Implications for Risk Stratification and Frontline Therapies in Pancreatic Cancer. *Comput. Struct. Biotechnol. J.* **2022**, *20*, 1154–1167. https://doi.org/10.1016/j.csbj.2022.02.031.
181. Principe, D.R.; Underwood, P.W.; Kumar, S.; Timbers, K.E.; Koch, R.M.; Trevino, J.G.; Munshi, H.G.; Rana, A. Loss of SMAD4 Is Associated With Poor Tumor Immunogenicity and Reduced PD-L1 Expression in Pancreatic Cancer. *Front Oncol.* **2022**, *12*, 806963. https://doi.org/10.3389/fonc.2022.806963.
182. Hosein, A.N.; Dougan, S.K.; Aguirre, A.J.; Maitra, A. Translational Advances in Pancreatic Ductal Adenocarcinoma Therapy. *Nat. Cancer* **2022**, *3*, 272–286. https://doi.org/10.1038/s43018-022-00349-2.
183. Qian, Y.; Gong, Y.; Fan, Z.; Luo, G.; Huang, Q.; Deng, S.; Cheng, H.; Jin, K.; Ni, Q.; Yu, X.; et al. Molecular Alterations and Targeted Therapy in Pancreatic Ductal Adenocarcinoma. *J. Hematol. Oncol.* **2020**, *13*, 130. https://doi.org/10.1186/s13045-020-00958-3.
184. Hessmann, E.; Johnsen, S.A.; Siveke, J.T.; Ellenrieder, V. Epigenetic Treatment of Pancreatic Cancer: Is There a Therapeutic Perspective on the Horizon? *Gut* **2017**, *66*, 168–179. https://doi.org/10.1136/gutjnl-2016-312539.
185. Baretti, M.; Ahuja, N.; Azad, N.S. Targeting the Epigenome of Pancreatic Cancer for Therapy: Challenges and Opportunities. *Ann. Pancreat. Cancer* **2019**, *2*, 18–18. https://doi.org/10.21037/apc.2019.10.01.
186. Yu, S.; Zhang, C.; Xie, K.-P. Therapeutic Resistance of Pancreatic Cancer: Roadmap to Its Reversal. *Biochim. Biophys. Acta (BBA)-Rev. Cancer* **2021**, *1875*, 188461. https://doi.org/10.1016/j.bbcan.2020.188461.
187. Bailey, P.; Chang, D.K.; Nones, K.; Johns, A.L.; Patch, A.-M.; Gingras, M.-C.; Miller, D.K.; Christ, A.N.; Bruxner, T.J.C.; Quinn, M.C.; et al. Genomic Analyses Identify Molecular Subtypes of Pancreatic Cancer. *Nature* **2016**, *531*, 47–52. https://doi.org/10.1038/nature16965.
188. Awasthi, N.; Kronenberger, D.; Stefaniak, A.; Hassan, M.S.; von Holzen, U.; Schwarz, M.A.; Schwarz, R.E. Dual Inhibition of the PI3K and MAPK Pathways Enhances Nab-Paclitaxel/Gemcitabine Chemotherapy Response in Preclinical Models of Pancreatic Cancer. *Cancer Lett.* **2019**, *459*, 41–49. https://doi.org/10.1016/j.canlet.2019.05.037.
189. Fell, J.B.; Fischer, J.P.; Baer, B.R.; Blake, J.F.; Bouhana, K.; Briere, D.M.; Brown, K.D.; Burgess, L.E.; Burns, A.C.; Burkard, M.R.; et al. Identification of the Clinical Development Candidate MRTX849, a Covalent KRAS G12C Inhibitor for the Treatment of Cancer. *J. Med. Chem.* **2020**, *63*, 6679–6693. https://doi.org/10.1021/acs.jmedchem.9b02052.
190. Janes, M.R.; Zhang, J.; Li, L.-S.; Hansen, R.; Peters, U.; Guo, X.; Chen, Y.; Babbar, A.; Firdaus, S.J.; Darjania, L.; et al. Targeting KRAS Mutant Cancers with a Covalent G12C-Specific Inhibitor. *Cell* **2018**, *172*, 578–589.e17. https://doi.org/10.1016/j.cell.2018.01.006.
191. Adamska, A.; Elaskalani, O.; Emmanouilidi, A.; Kim, M.; Abdol Razak, N.B.; Metharom, P.; Falasca, M. Molecular and Cellular Mechanisms of Chemoresistance in Pancreatic Cancer. *Adv. Biol. Regul.* **2018**, *68*, 77–87. https://doi.org/10.1016/j.jbior.2017.11.007.
192. Awad, M.M.; Liu, S.; Rybkin, I.I.; Arbour, K.C.; Dilly, J.; Zhu, V.W.; Johnson, M.L.; Heist, R.S.; Patil, T.; Riely, G.J.; et al. Acquired Resistance to KRAS G12C Inhibition in Cancer. *N. Engl. J. Med.* **2021**, *384*, 2382–2393. https://doi.org/10.1056/NEJMoa2105281.
193. Lito, P.; Solomon, M.; Li, L.-S.; Hansen, R.; Rosen, N. Allele-Specific Inhibitors Inactivate Mutant KRAS G12C by a Trapping Mechanism. *Science* **2016**, *351*, 604–608. https://doi.org/10.1126/science.aad6204.
194. Gurtner, K.; Kryzmien, Z.; Koi, L.; Wang, M.; Benes, C.H.; Hering, S.; Willers, H.; Baumann, M.; Krause, M. Radioresistance of KRAS/TP53-mutated Lung Cancer Can Be Overcome by Radiation Dose Escalation or EGFR Tyrosine Kinase Inhibition in Vivo. *Int. J. Cancer* **2020**, *147*, 472–477. https://doi.org/10.1002/ijc.32598.
195. Huang, B.; Wang, J.; Chen, Q.; Qu, C.; Zhang, J.; Chen, E.; Zhang, Y.; Wang, Y.; Ni, L.; Liang, T. Gemcitabine Enhances OSI-027 Cytotoxicity by Upregulation of MiR-663a in Pancreatic Ductal Adenocarcinoma Cells. *Am. J. Transl. Res.* **2019**, *11*, 473–485.
196. Huang, L.; Hu, C.; Cao, H.; Wu, X.; Wang, R.; Lu, H.; Li, H.; Chen, H. MicroRNA-29c Increases the Chemosensitivity of Pancreatic Cancer Cells by Inhibiting USP22 Mediated Autophagy. *Cell. Physiol. Biochem.* **2018**, *47*, 747–758. https://doi.org/10.1159/000490027.
197. Fan, P.; Liu, L.; Yin, Y.; Zhao, Z.; Zhang, Y.; Amponsah, P.S.; Xiao, X.; Bauer, N.; Abukiwan, A.; Nwaeburu, C.C.; et al. MicroRNA-101-3p Reverses Gemcitabine Resistance by Inhibition of Ribonucleotide Reductase M1 in Pancreatic Cancer. *Cancer Lett.* **2016**, *373*, 130–137. https://doi.org/10.1016/j.canlet.2016.01.038.

Pobrano z / Downloaded from Repository of Wrocaw University of Environmental and Life Sciences 2025-10-15 21:43

198. Li, C.; Zhao, Z.; Zhou, Z.; Liu, R. Linc-ROR Confers Gemcitabine Resistance to Pancreatic Cancer Cells via Inducing Autophagy and Modulating the MiR-124/PTBP1/PKM2 Axis. *Cancer Chemother. Pharmacol.* **2016**, *78*, 1199–1207. https://doi.org/10.1007/s00280-016-3178-4.
199. REN, Z.-G.; DONG, S.-X.; HAN, P.; QI, J. MiR-203 Promotes Proliferation, Migration and Invasion by Degrading SIK1 in Pancreatic Cancer. *Oncol. Rep.* **2016**, *35*, 1365–1374. https://doi.org/10.3892/or.2015.4534.
200. Gu, J.; Wang, D.; Zhang, J.; Zhu, Y.; Li, Y.; Chen, H.; Shi, M.; Wang, X.; Shen, B.; Deng, X.; et al. GFRα2 Prompts Cell Growth and Chemoresistance through Down-Regulating Tumor Suppressor Gene PTEN via Mir-17-5p in Pancreatic Cancer. *Cancer Lett.* **2016**, *380*, 434–441. https://doi.org/10.1016/j.canlet.2016.06.016.
201. Tian, X.; Shivapurkar, N.; Wu, Z.; Hwang, J.J.; Pishvaian, M.J.; Weiner, L.M.; Ley, L.; Zhou, D.; Zhi, X.; Wellstein, A.; et al. Circulating MicroRNA Profile Predicts Disease Progression in Patients Receiving Second-Line Treatment of Lapatinib and Capecitabine for Metastatic Pancreatic Cancer. *Oncol Lett.* **2016**, *11*, 1645–1650. https://doi.org/10.3892/ol.2016.4101.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Pobrano z / Downloaded from Repository of Wrocaw University of Environmental and Life Sciences 2025-10-15 21:43

HHS Public Access
Author manuscript
Nat Genet. Author manuscript; available in PMC 2017 August 16.

Published in final edited form as:
Nat Genet. 2017 May ; 49(5): 789–794. doi:10.1038/ng.3823.

Genome-wide association study of glioma subtypes identifies specific differences in genetic susceptibility to glioblastoma and non-glioblastoma tumors

A full list of authors and affiliations appears at the end of the article.

Abstract

Genome-wide association studies (GWAS) have transformed our understanding of glioma susceptibility, but individual studies have had limited power to identify risk loci. We performed a meta-analysis of existing GWAS and two new GWAS, which totaled 12,496 cases and 18,190 controls. We identified five new loci for glioblastoma (GBM) at 1p31.3 (rs12752552; $P = 2.04 \times 10^{-9}$, odds ratio (OR) = 1.22), 11q14.1 (rs11233250; $P = 9.95 \times 10^{-10}$, OR = 1.24), 16p13.3 (rs2562152; $P = 1.93 \times 10^{-8}$, OR = 1.21), 16q12.1 (rs10852606; $P = 1.29 \times 10^{-11}$, OR = 1.18) and 22q13.1 (rs2235573; $P = 1.76 \times 10^{-10}$, OR = 1.15), as well as eight loci for non-GBM tumors at 1q32.1 (rs4252707; $P = 3.34 \times 10^{-9}$, OR = 1.19), 1q44 (rs12076373; $P = 2.63 \times 10^{-10}$, OR = 1.23), 2q33.3 (rs7572263; $P = 2.18 \times 10^{-10}$, OR = 1.20), 3p14.1 (rs11706832; $P = 7.66 \times 10^{-9}$, OR = 1.15), 10q24.33 (rs11598018; $P = 3.39 \times 10^{-8}$, OR = 1.14), 11q21 (rs7107785; $P = 3.87 \times 10^{-10}$, OR = 1.16), 14q12 (rs10131032; $P = 5.07 \times 10^{-11}$, OR = 1.33) and 16p13.3 (rs3751667; $P = 2.61 \times 10^{-9}$, OR = 1.18). These data substantiate that genetic susceptibility to GBM and non-GBM tumors are highly distinct, which likely reflects different etiology.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

Correspondence should be addressed to B.S.M. (beatrice.melin@umu.se), R.S.H. (richard.houlston@icr.ac.uk) or M.L.B. (mbondy@bcm.edu).

[25] A full list of members and affiliations appears in the Acknowledgments

[41] These authors contributed equally to this work

[42] These authors jointly directed this work

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

AUTHOR CONTRIBUTIONS

M.L.B., B.S.M., R.S.H. and J.S.B.-S. managed the project; R.S.H., M.L.B., B.S.M., J.S.B.-S., R.B.J., Q.T.O., B.K. and M.R.W. drafted the manuscript; Q.T.O., K.L., B.K., J.E.E.-P. and P.A.D. performed statistical analyses; Y.C., K.L., Y.L. and B.K. performed bioinformatics analyses; B.S.M., J.S.B.-S., M.R.W., J.K.W., C.J., D.I., R.K.L., G.A., P.A.D., U.A., T.R., H.H., L.M., M.L.K., H.S., J.L.B., F.D., D.L, C.I.A, C.L., R.T.M., J. Shildkraut, F.A.-O., S. Sadetski, M. Scheurer, S. Shete, E.B.C., S.H.O., R.B.J., R.S.H. and M.L.B. developed the GICC protocol and performed sample acquisition; and P.R., S.C., M. Linet, Z.W. and M.Y. provided the National Cancer Institute (NCI) data. In the UK, P.B., A.S., M.J.S., S.J.F. and R.S.H. developed patient recruitment, performed sample acquisition and performed sample collection of cases; P.B. oversaw DNA isolation and storage, and performed case and control ascertainment, and supervision of DNA extractions. In Germany, M. Simon, M.M.N., H.-E.W., S. Schreiber and J. Schramm developed patient recruitment, and oversaw performed blood sample collection; M. Simon oversaw DNA isolation and storage and performed case and control ascertainment, and supervision of DNA extractions; and S. Herms, S. Heilmann and K.G. performed experimental work. In France, M. Sanson and J.-Y.D. developed patient recruitment; M. Labussière, A.-L.D.S., P.G., K.M., A.I. and K.H.-X. performed patient ascertainment. M. Lathrop performed laboratory management and oversaw genotyping of the French samples. All authors contributed to the final manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Glioma accounts for around 27% of all primary brain tumors and is responsible for approximately 13,000 cancer-related deaths in the United States each year[1, 2]. Gliomas can be broadly classified into GBM and lower-grade non-GBM tumors[3]. Gliomas typically have a poor prognosis irrespective of medical care, with the most common form, GBM, having a five-year survival rate of only 5% (ref. 4).

So far, no environmental exposures have been robustly linked to the risk of developing glioma, except for moderate to high doses of ionizing radiation, which accounts for a small proportion of cases[5]. Evidence for an inherited predisposition to glioma is provided by a number of rare inherited cancer syndromes, such as Turcot's and Li–Fraumeni syndromes, as well as neurofibromatosis. Even collec-tively, however, these account for little of the twofold familial risk of glioma[6]. Our understanding of the heritability of glioma has been transformed by recent GWAS, which have identified single-nucleotide polymorphisms (SNPs) at 13 loci influencing risk[7–14].

Previous individual studies have had limited statistical power for the additional discovery of new glioma risk loci[15]. Therefore, to gain more comprehensive insight into glioma etiology, we performed a meta-analysis of previously published GWAS and two new GWAS, which allowed us to identify 13 new risk loci for glioma.

We analyzed GWAS SNP data that passed quality control for 12,496 cases (6,191 classified as GBM and 5,819 classified as non-GBM tumors) and 18,190 controls from eight studies with individuals of European ancestry, a new GWAS of 4,572 cases and 3,286 controls performed by the Glioma International Case Control Consortium (GICC) (Supplementary Table 1), a new GWAS of 1,591 cases and 804 controls from the University of California, San Francisco (UCSF)-Mayo, and six previously reported GWAS[9, 10, 13] totaling 6,405 cases and 14,100 controls (Supplementary Table 2). To increase genomic resolution, we imputed >10 million SNPs. Quantile–quantile (Q-Q) plots for SNPs with a minor allele frequency (MAF) >1% after imputation did not show evidence of substantive overdispersion (λ = 1.02–1.10, λ_{90} = 1.02–1.05; Supplementary Fig. 1). We derived joint ORs and 95% confidence intervals (CIs) under a fixed-effects model for each SNP with MAF >1% and associated per-allele principal component (PCA) corrected P-values for all glioma, GBM and non-GBM cases versus those for the controls (Fig. 1).

In the combined meta-analysis, among previously published glioma risk SNPs, those for all glioma at 17p13.1 (*TP53*), for GBM at 5p15.33 (*TERT*), 7p11.2 (*EGFR*), 9p21.3 (*CDKN2B–AS1*) and 20q13.33 (*RTEL1*), and for non-GBM tumors at 8q24.21 (*CCDC26*), 11q23.2, 11q23.3 (*PHLDB1*) and 15q24.2 (*ETFA*) showed even greater evidence for association (Supplementary Fig. 2 and Supplementary Table 3). SNPs at 10q25.2 and 12q12.1 for non-GBM tumors retained genome-wide significance (i.e., $P < 5.0 \times 10^{-8}$). Associations at the previously reported 3q26.2 (near *TERC*)[11] and 12q23.33 (*POLR3B*)[10] loci for GBM did not retain statistical significance (P values for the most associated SNPs are 2.68×10^{-5} and 1.60×10^{-5}, respectively; Supplementary Table 3).

In addition to previously reported loci, we identified genome-wide significant associations marking new risk loci (Table 1, Supplementary Fig. 3 and Supplementary Data 1) for GBM

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

at 1p31.3 (rs12752552; $P = 2.04 \times 10^{-9}$), 11q14.1 (rs11233250; $P = 9.95 \times 10^{-10}$), 16p13.3 (rs2562152; $P = 1.93 \times 10^{-8}$), 16q12.1 (rs10852606; $P = 1.29 \times 10^{-11}$) and 22q13.1 (rs2235573; $P = 1.76 \times 10^{-10}$) and for non-GBM tumors at 1q32.1 (rs4252707; $P = 3.34 \times 10^{-9}$), 1q44 (rs12076373; $P = 2.63 \times 10^{-10}$), 2q33.3 (rs7572263; $P = 2.18 \times 10^{-10}$), 3p14.1 (rs11706832; $P = 7.66 \times 10^{-9}$), 10q24.33 (rs11598018; $P = 3.39 \times 10^{-8}$), 11q21 (rs7107785; $P = 3.87 \times 10^{-10}$), 14q12 (rs10131032; $P = 5.07 \times 10^{-11}$) and 16p13.3 (rs3751667; $P = 2.61 \times 10^{-9}$). Conditional analysis confirmed the existence of two independent association signals at 7p11.2 (*EGFR*) as previously reported[7] but did not provide evidence for additional signals at any of the other established identified risk loci or at the 13 newly identified loci. Case-only analyses con-firmed the specificity of 11q14.1, 16p13.3 and 22q13.1 associations for GBM and of 1q44, 2q33.3, 3p14.1, 11q21 and 14q12 associations for non-GBM tumors (Fig. 2 and Supplementary Table 4). Collectively, our findings provide strong evidence for specific associations for the different glioma subtypes, consistent with their previously described distinctive molecular profiles, presumably resulting from different etiological pathways.

Across the new and known risk loci, we found a significant enrichment of overlap with enhancers in H9-Derived neuronal progenitor cells ($P = 8.2 \times 10^{-5}$; Supplementary Data 2). These observations support the assertion that the loci identified in the GWAS influence glioma risk through effects on neural *cis* regulatory networks and that they are strongly involved in transcriptional initiation and enhancement. To gain further insight into the biological basis for associations at the 13 new risk loci, we performed an expression quantitative trait loci (eQTL) analysis using RNA-seq data on ten regions of normal human brain from up to 103 individuals from the Genotype–Tissue Expression (GTEx) project[16] and blood eQTL data on 5,311 individuals from Westra *et al.*[17]. We used summary-level mendelian randomization (SMR)[18] analysis to test for a concordance between signals from GWAS and *cis* eQTL for genes within 1 Mb of the sentinel and correlated SNPs ($r^2 > 0.8$) at each locus (Supplementary Data 3) and derived b_{XY} statistics, which estimate the effect of gene expression on glioma risk. Additionally, for each of the risk SNPs at the 13 new loci (as well as the correlated variants), we examined published data[19, 20] and made use of the online resources HaploRegv4, RegulomeDB and SeattleSeq for evidence of functional effects (Supplementary Table 5).

At 16q12.1, the GBM association signal was significantly associated with *HEATR3* expression in nine of ten regions of the brain (PSMR = 3.38×10^{-6} to 6.55×10^{-10}; $b_{XY} = 0.14$–0.24; Supplementary Fig. 4 and Supplementary Data 3). The risk allele ‘C’ of rs10852606 that was associated with reduced *HEATR3* expression was consistent with differential expression of *HEATR3* being the functional basis of the 16q12.1 association. The observation that variation at 16q12.1 is associated with risk of testicular[21] (rs8046148; pairwise r^2 and D' with rs10852606 of 0.67 and 1.0, respectively) and esophageal[22] (rs4785204; pairwise r^2 and D' with rs10852606 of 0.16 and 1.0, respectively) cancer suggests that the locus has pleiotropic effects on tumor risk, which are compatible with generic effects as shown by the observation of a *HEATR3* eQTL signal in blood (*PSMR* = 5.84×10^{-11}; $b_{XY} = 0.30$).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Similarly, significant associations between gene expression and glioma risk were observed at the GBM loci 1p31.3 (*JAK1*, brain cortex and cerebellar hemisphere), 16p13.3 (*POLR3K*, whole blood) and 22q13.1 (*CTA-228A9.3*, brain cerebellum; *PICK1*, brain hippocampus) (Supplementary Fig. 4 and Supplementary Data 3). The non-GBM association at 1q32.1 marked by rs4252707 (Supplementary Fig. 3) maps to intron 8 of the gene encoding MDM4, a p53-binding protein. The SNP rs4252707 is in strong linkage disequilibrium (LD) with rs12031912 and rs12028476 ($r^2 = 0.92$), both of which map to the *MDM4* promoter. Although no significant eQTL was shown in any brain tissue, an association with *MDM4* was seen in blood (PSMR = 4.74 × 10^{-6}; $b_{XY} = 0.31$; Supplementary Fig. 4 and Supplementary Data 3). Overexpression of *MDM4* is a feature in glioma tumors containing wild-type *TP53* and no amplification of the *MDM2* gene, consistent with *MDM4* amplification being a mechanism by which the p53-dependent growth control is inactivated[23].

The 1q44 association with non-GBM that is marked by rs12076373 maps to intron 8 of *AKT3*, whose encoded product is one of the major downstream effectors of phosphatidylinositol 3-kinase (PI3K) and is highly expressed during active neurogenesis, with haploinsufficiency causing postnatal microcephaly and agenesis of the corpus callosum[24]. Notably, *AKT3* is hyper-expressed in glioma, thus having a role in tumor viability by activating DNA repair[25]. Although rs12076373 does not map to a regulatory element, the correlated SNPs rs12124113 ($r^2 = 0.94$) and rs59953491 ($r^2 = 0.90$) locate within an enhancer element in brain cells and tissues, including H9-derived neuronal progenitor cultured cells, cortex-derived primary cultured neurospheres and NH-A astrocytes.

The 3p14.1 association with non-GBM that is marked by rs11706832 localizes to intron 2 of *LRIG1*. Although we did not identify an eQTL in this gene, *LRIG1* is highly expressed in the brain and is a pan-negative regulator of the epidermal growth factor receptor (EGFR) signaling pathway, which inhibits hypoxia-induced vasculogenic mimicry via EGFR–PI3K–AKT pathway suppression and epithelial-to-mesenchymal transition[26]. Reduced *LRIG1* expression is linked to tumor aggressiveness, temozolomide resistance and radio-resistance[27, 28]. We have previously shown an association for glioma at *EGFR* (7p11.2)[7], which is well established to be pivotal in both the initiation of primary GBM and the progression of lower-grade glioma to grade IV. Although speculative, our new findings now suggest a more extensive pathway involving variation at *LRIG1* and *AKT3*.

Of particular interest is rs7572263, which maps to 2q33.3, localizes within intron 3 of *C2orf80* and is 50 kb telomeric to *IDH1*. Mutation of *IDH1* is a driver for gliomagenesis[29, 30] and is responsible for the CpG island methylator (G-CIMP) phenotype[31, 32]. Mutations in *IDH1* predominate in non-GBM glioma[33, 34]; therefore, the association at 2q33.3 is plausible as a basis for susceptibility to non-GBM glioma. In the absence of convincing eQTL or other functional support, this does not preclude *C2orf80* or another gene mapping to the region of LD as being the functional basis for the 2q33.3 association.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

The maintenance of telomeres is central to cell immortalization, and it has a central role in gliomagenesis[35]. We have previously shown that the risk of GBM is strongly linked to genetic variation in the telomere-related genes *TERT* (5p15.33) and *RTEL1* (20q13.33), and possibly also *TERC* (3q26.2)[8, 9, 11]. The 10q24.33 association with non-GBM that is marked by rs11598018 lies intronic to *OBFC1*, which functions in a telomere-associated complex that protects telomeres independently of POT1 (ref. 36). The CST complex, whose components are encoded by *OBFC1, CTC1*, and *TEN1*, competes with shelterin for telomeric-DNA-inhibiting telomerase-based telomere extension[37]. The significant association between the risk of non-GBM tumors and *OBFC1* variation is particularly of note in light of our recent exome-sequencing report demonstrating that rare germline loss-of-function mutations in genes that encode components of the shelterin complex are a cause of familial oligodendroglioma[38]. The glioma risk alleles at *TERT, TERC* and *OBFC1* are associated with increased leukocyte telomere length, thereby supporting a relationship between genotype and biology (Supplementary Table 6)[35, 39, 40]. However, the *RTEL1* locus is not consistent with such a postulate, and recent data that have not shown a relationship between mutations in the *TERT* promoter and telomere length in glioma[41] raise the possibility of a role for extratelomeric effects.

The deregulation of pathways involved in telomere length and EGFR signaling are thus consistent with glioma risk being governed by pathways that are important in the longevity of glial cells, and they substantiate early observations that genetic susceptibility to GBM and non-GBM tumors is highly distinct, presumably reflecting different etiologies between GBM and non-GBM tumors (Fig. 2).

The other associations we identified mark genes with varying degrees of plausibility for having a role in glioma oncogenesis. The GBM association at 16p13.33 marked by rs2562152 localizes 3 kb telomeric to *MPG*, which encodes a *N*-methylpurine DNA glycosylase whose expression is linked to temozolomide resistance in glioma[42]. Although attractive as a candidate, the only genes for which there was found to be a significant association between expression and glioma risk were *POLR3K* and *C16ORF33* in blood (Supplementary Fig. 4 and Supplementary Data 3). At 1p31.3, only *JAK1* provided convincing evidence for a significant eQTL with glioma risk SNPs in brain tissue. The strongest association was shown in the cortex ($P_{SMR} = 1.61 \times 10^{-6}$; $b_{XY} = 0.22$; Supplementary Fig. 4 and Supplementary Data 3), with the risk allele 'T' of rs12752552 showing increased *JAK1* expression. The *cis*-eQTL signal for *JAK1* in the cortex maps from 65.3 Mb to 65.35 Mb and shows a consistent direction of effect with the glioma-associated SNPs. JAK1–STAT6 signaling is increasingly being recognized to be relevant in glioma progression[43]. Hence, although *JAK1* remains an attractive candidate mechanistic basis for the glioma association at 1p31.3, we cannot exclude the possibility that the cluster of SNPs between 65.3 Mb and 65.35 Mb contains the true causal variant. In the absence of functional data, potential target genes for associations at 11q14.1 (GBM), 16p13.3 (non-GBM), 11q21 (non-GBM) and 14q12 (non-GBM) remain to be elucidated.

In conclusion, we have performed the largest glioma GWAS to date and have identified 13 new glioma risk loci, thereby providing further evidence for a polygenic basis of genetic susceptibility to glioma. Histological classification of glioma is, in part, being superseded by

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

molecular profiling[34, 44]; hence, it is important to understand the biology behind these risk variants in the context of molecularly defined glioma subtypes. Currently identified risk SNPs for glioma account for, at best, ~27% and ~37% of the familial risk of GBM and non-GBM tumors, respectively (Supplementary Table 7). Therefore, further GWAS-based analyses in concert with functional analyses should lead to additional insights into the biology and etiological basis of the different glioma histologies. Notably, such information can inform gene discovery initiatives and thus have a measurable effect on the successful development of new therapeutic agents.

ONLINE METHODS

Ethics

Collection of patient samples and associated clinico-pathological information was undertaken with written informed consent and relevant ethical review board approval at the respective study centers in accordance with the tenets of the Declaration of Helsinki. Specifically informed consent and ethical board approval was obtained from the South-East Multicentre Research Ethics Committee (MREC) (UK), the Scottish MREC (UK), the APHP ethical committee-CPP (Comité de Protection des Personnes) (France), the Ethics Commission of the Medical Faculty of the University of Bonn (Germany), the University of Texas MD Anderson Cancer Institutional Review Board (USA), the Mayo Clinic Office for Human Research Protection (USA), the UCSF Committee on Human Research (USA), the University Hospitals of Cleveland Institutional Review Board (USA) and the Cleveland Clinic Institutional Review Board (board for the Case Comprehensive Cancer Center) (USA). The diagnosis of glioma (ICDO-3 codes 9380-9480 or equivalent) was established through histology in all cases in accordance with World Health Organization guidelines. Every effort was made to classify tumors as GBM or non-GBM.

GWAS data sets

GICC, UK, French, German, MDA, SFAGS and GliomaScan—Studies participating in GICC are described in Amirian *et al.*[46] and in Supplementary Table 1. Briefly, they comprise 5,189 glioma cases and 3,827 controls that were ascertained through centers in the USA, Denmark, Sweden and the UK. Cases had newly diagnosed glioma, and controls had no personal history of central nervous system tumor at the time of ascertainment. Detailed information regarding recruitment protocol is given in Amirian *et al.*[46]. Cases and controls were genotyped using the Illumina Oncoarray according to the manufacturer's recommendations (Illumina Inc.). Individuals with a call rate <99%, as well as all individuals evaluated to be of non-European ancestry (<80% estimated European ancestry using the FastPop[47] procedure developed by the GAMEON consortium with HapMap version 2 CEU, JPT/CHB and YRI populations as a reference; Supplementary Fig. 5), were excluded. For pairs of apparent first-degree relatives, we removed the control from a case–control pair; otherwise, we excluded the individual with the lower call rate. SNPs with a call rate <95% were excluded as were those with a MAF <0.01 or those displaying significant deviation from the Hardy–Weinberg equilibrium (HWE) (i.e., $P < 10^{-5}$). After performing these quality-control measures, there were 4,572 cases and 3,286 controls remaining for downstream analyses.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

The UK, French, German, MDA, SFAGS and GliomaScan GWAS of non-overlapping case–control series of Northern European ancestry have been the subject of previous studies. Briefly, the UK GWAS[7, 8, 10] was based on 636 cases (401 males; mean age 46 years) who were ascertained through the INTERPHONE study[48]. Individuals from the 1958 Birth Cohort ($n = 2{,}930$) served as a source of controls. The French GWAS[7, 10] comprised 1,495 patients with glioma who were ascertained through the Service de Neurologie Mazarin, Groupe Hospitalier Pitié-Salpêtrière Paris. The controls ($n = 1{,}213$) were ascertained from the SU.VI.MAX (Supplementation en Vitamines et MinerauxAntioXydants) study of 12,735 healthy subjects (women aged 35–60 years; men aged 45–60 years)[49]. The German GWAS[10] comprised 880 patients who had undergone surgery for a glioma at the Department of Neurosurgery, University of Bonn Medical Center, between 1996 and 2008. Control subjects were taken from three population studies: KORA (Co-operative Health Research in the Region of Augsburg; $n = 488$)[50]; POPGEN (Population Genetic Cohort; $n = 678$)[51] and the Heinz Nixdorf Recall study ($n = 380$)[52]. Standard quality-control measures were applied to the UK, French and German GWAS and have previously been reported. The MDA GWAS[8] was based on 1,281 cases (786 males; mean age 47 years) who were ascertained through the MD Anderson Cancer Center, Texas, between 1990 and 2008. Individuals from the Cancer Genetic Markers of Susceptibility (CGEMS, $n = 2{,}245$) studies served as controls[53, 54]. Quality-control measures were applied as per the primary GWAS. The UCSF adult glioma case–control study (SFAGS–GWAS) included participants of the San Francisco Bay Area Adult Glioma Study (AGS). Details of subject recruitment for AGS have been reported previously[9, 12, 34, 55, 56]. Briefly, cases were adults (>18 years of age) with newly diagnosed, histologically confirmed glioma. Population-based cases who were diagnosed between 1991 and 2009 (series 1–4) and who were residing in the six San Francisco Bay area counties were ascertained using the Cancer Prevention Institute of California's early-case ascertainment system. Clinic-based cases who were diagnosed between 2002 and 2012 (series 3–5) were recruited from the UCSF Neuro-oncology Clinic, regardless of the place of residence. From 1991 to 2010, population-based controls from the same residential area as the population-based cases were identified using random digit-dialing and were frequency matched to population-based cases for age, gender and ethnicity. Between 2010 and 2012, all controls were selected from the UCSF general medicine phlebotomy clinic. Clinic-based controls were matched to clinic-based glioma cases for age, gender and ethnicity. Consenting participants provided blood, buccal and/or saliva specimens, and information, during in-person or telephone interviews. A total of 677 cases and 3,940 controls (including 3,347 Illumina iControlDB iControls) were used in the current analysis. For the GliomaScan GWAS[13], in addition to the published analysis, we excluded samples from the ATBC (Finnish study) and controls from NSHDS due to exhibiting outlying population ancestry after manual inspection of PCA plots. In total 1,653 cases and 2,725 controls were used in the current study.

GWAS data from the seven studies were imputed to >10 million SNPs with IMPUTE2 (v2.3)[57] software using a merged reference panel consisting of data from the 1000 Genomes Project (phase 1 integrated release 3, March 2012)[58] and UK10K (ALSPAC, EGAS00001000090 and EGAD00001000195, and TwinsUK EGAS00001000108 and EGAS00001000194 studies). Genotypes were aligned to the positive strand in both

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

imputation and genotyping. Imputation was conducted separately for each study, and in each the data were pruned to a common set of SNPs between cases and controls before imputation. We set thresholds for imputation quality to retain potential risk variants with MAF > 0.01. Poorly imputed SNPs, defined by an information measure <0.40 with IMPUTE2, were excluded, as were SNPs exhibiting a significant deviation from Hardy–Weinberg equilibrium ($P < 1 \times 10^{-8}$) in controls. Test of association between imputed SNPs and glioma was performed using SNPTEST (v2.5)[59] under an additive frequentist model. The adequacy of the case–control matching and the possibility of differential genotyping of cases and controls were formally evaluated using Q-Q plots of test statistics (Supplementary Fig. 1). Where appropriate, principal components, generated using common SNPs, were included in the analysis to limit the effects of cryptic population stratification that otherwise might cause inflation of test statistics. Principal components, based on genotyped SNPs, were generated for the GICC, GliomaScan, MDA-GWAS and SFAGS studies using PLINK[60]. Eigenvectors for the German GWAS were inferred using smartpca (part of EIGENSOFTv2.4)[61] by merging cases and controls with Phase II HapMap samples[10]. PCA plots for all studies are provided in Supplementary Figure 4.

UCSF-Mayo GWAS—The UCSF-Mayo study comprised Mayo cases ($n = 945$) and UCSF cases ($n = 574$) and Mayo Clinic Biobank control ($n = 806$) data. The Mayo Clinic case–control study has been described previously[9, 34, 62]. Briefly, adult cases (>18 years of age) were identified at diagnosis (diagnosed at Mayo Clinic) or at pathologic confirmation (diagnosed elsewhere and treated at Mayo Clinic), and the patients had a surgical resection or biopsy between 1973 and 2014. Consenting participants provided blood, buccal and/or saliva specimens, and information, during in-person or telephone interviews. This analysis used 574 non-overlapping cases from the UCSF Adult Glioma Study described above. Mayo Clinic and UCSF cases were genotyped using the Illumina Oncoarray. The Mayo Clinic Biobank controls comprised volunteers who donated biological specimens and provided risk factor data, access to clinical data obtained from the medical record and consent to participate in any study approved by the Access Committee. Recruitment for the Mayo Clinic Biobank took place from April 2009 through December 2015. Although participants could be unselected volunteers, the vast majority of participants were contacted as part of a pre-scheduled medical examination in the Department of Medicine, Divisions of Community Internal Medicine, Family Medicine and General Internal Medicine at Mayo Clinic sites in Rochester (Minnesota), Jacksonville (Florida), and the Mayo Clinic Health System sites in La Crosse and Onalaska (Wisconsin). All individuals were aged 18 years and older at the time of consent. Illumina Omni Express genotyping arrays were run on the 806 Mayo Clinic Biobank participants.

Quality-control analyses were performed on each cohort separately (Mayo cases, UCSF cases and Mayo Clinic Biobank controls). SNPs with call rates <95% were removed, followed by removal of subjects with call rates <95%. Concordance of replicate samples was assessed, and the sample with the higher call rate was retained. Subject's sex was verified using the sex check option in PLINK. Relationship checking was performed by estimating the proportion of alleles shared identical by descent (IBD) for all pairs of subjects in PLINK[60]. STRUCTURE[63] was used to assess population admixture with 1000 Genomes as

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

a reference. Subjects indicated to be non-Caucasian were excluded. Prior to imputation, SNPs were tested for HWE, and SNPs with HWE $P < 10^{-6}$ were removed. Mayo Clinic, UCSF and Mayo Clinic Biobank SNP data were each phased and imputed using the Michigan Imputation Server with the Haplotype Reference Consortium (release 1; http://www.haplotype-reference-consortium.org) as reference. Genotypes were forward-strand-aligned to the 1000 Genomes reference, and for ambiguous SNPs the Browning strand checking utility was used (http://faculty.washington.edu/sguy/beagle/strand_switching/strand_switching.html). PCA was used to correct for population stratification using SNPs common to cases and controls. The first three principal components were significantly ($P <$ 0.05) associated with case–control status. An additive logistic regression model was used to assess the association between each SNP and disease status, with genotype being coded as 0, 1 or 2 copies of the minor allele, adjusted for age, sex and the first three principal components.

Meta-analysis and additional statistical analyses

Meta-analyses were performed using the fixed-effects inverse-variance method based on the β-esti-mates and standard errors from each study using META (v1.6)[64]. Cochran's Q-statistic was used to test for heterogeneity, and the I^2 statistic was used to quantify the proportion of the total variation due to heterogeneity[65], taking I^2 values >75 to indicate significant heterogeneity. Using the meta-analysis summary statistics and LD correlations from a reference panel of the 1000 Genomes Project combined with UK10K, we used GCTA[66, 67] to perform conditional association analysis. Association statistics were calculated for all SNPs, conditioning on the top SNP in each locus showing genome-wide significance. This was carried out in a step-wise fashion. We performed a case-only analysis to test for differences in SNP-risk-allele frequency between GBM and non-GBM tumors.

ENCODE and chromatin state dynamics

Risk SNPs and their proxies (i.e., $r^2 > 0.8$ in the 1000 Genomes EUR reference panel) were annotated for putative functional effect using HaploReg (v4)[68], RegulomeDB[69] and SeattleSeq Annotation[70]. These servers make use of data from ENCODE, genomic evolutionary rate profiling (GERP) conservation metrics, combined annotation-dependent depletion (CADD) scores and PolyPhen scores. We searched for overlap of associated SNPs with enhancers defined by the FANTOM5 enhancer atlas[19], annotating by overlap with ubiquitous, permissive and robust enhancers, as well as enhancer–promoter correlations and enhancers specifically expressed in astrocytes, neuronal stem cells and brain tissue. Similarly, we searched for overlap with 'super-enhancer' regions, as defined by Hnisz *et al.*[20], restricting analysis to data from U87 GBM cells, astrocyte cells and brain tissue. We additionally made use of 15-state chromHMM data from H1- and H9-derived neuronal progenitor cells available from the Epigenome Roadmap Project[71]. Enhancer enrichment analysis was carried out using HaploReg (v4.0)[68]. Briefly, from a query list of variants, the overlap with enhancers in each of 107 cell types, as predicted from the Roadmap Epigenomics Project[71] chromatin-state segmentations, was calculated. A binomial test for enrichment was performed against a background set of all (i) 1000 Genomes variants with MAF > 0.05 and (ii) all unique GWAS loci in the European population. We applied a cutoff of $P < 3.94 \times 10^{-4}$ corresponding to a Bonferroni correction for 127 cell lines and tissues.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Expression quantitative trait loci (eQTL) analysis

To examine the relationship between SNP genotype and gene expression, we carried out summary-data-based mendelian randomization (SMR) analysis as per Zhu *et al.*[18] (at http://cnsgenomics.com/software/smr/index.html). We used publicly available brain tissue data from the GTEx[16] (http://www.gtexportal.org) v6p release. Briefly, GWAS summary statistics files were generated from the meta-analysis. Reference files were generated from merging 1000 Genomes phase 3 and UK10K (ALSPAC and TwinsUK) vcfs. Summary eQTL files for GTEx samples were generated from downloaded v6p "all_snpgene_pairs" files. Besd files were generated from these summary eQTL files using the –make-besd command. Additionally, we analyzed downloaded whole-blood eQTL data from Westra *et al.*[17]. Results from the SMR test for each of the 13 new glioma loci are reported in Supplementary Data 3. As previously advocated[18], only probes with at least one eQTL P value $<5.0 \times 10^{-8}$ were considered for SMR analysis. We set a threshold for the SMR test of $P_{SMR} < 1.06 \times 10^{-4}$ corresponding to a Bonferroni correction for 473 tests (473 probes with a top eQTL $P < 5.0 \times 10^{-8}$ across the 13 loci, 10 brain regions and Westra data set). For all genes passing this threshold, we generated plots of the eQTL and GWAS associations at the locus, as well as plots of GWAS and eQTL effect sizes (i.e., corresponding to input for the HEIDI heterogeneity test). HEIDI test P values <0.05 were taken to indicate significant heterogeneity. Respective SMR plots for significant eQTLs are shown in Supplementary Figure 4.

Additional statistical and bioinformatics analysis

Estimates of individual variance in risk associated with glioma risk SNPs was carried out using the method described in Pharoah *et al.*[72], assuming the familial risk of high-grade and low-grade glioma to be 1.76 and 1.54, respectively, from analysis of the Swedish series in Scheurer *et al.*[73]. Briefly, for a single allele (i) of frequency p, relative risk R and ln risk r, the variance (V_i) of the risk distribution due to that allele is given by:

$$V_i=(1-p)^2E^2+2p(1-p)(r-E)^2+p^2(2r-E)^2$$

Where E is the expected value of r given by:

$$E=2p(1-p)r+2p^2r$$

For multiple risk alleles, the distribution of risk in the population tends toward the normal with variance:

$$V=\sum V_i$$

The total genetic variance (V) for all susceptibility alleles has been estimated to be √1.77. Thus, the fraction of the genetic risk explained by a single allele is given by:

$$V_i/V$$

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

LD metrics were calculated in vcftools (v0.1.12b)[74] using UK10K data and plotted using visPIG[75]. LD blocks were defined on the basis of HapMap recombination rate (cM/Mb), as defined using the Oxford recombination hotspots and on the basis of distribution of confidence intervals defined by Gabriel *et al.*[76].

Data availability

Genotype data from the GICC GWAS are available from the database of Genotypes and Phenotypes (dbGaP) under accession phs001319.v1.p1. Additionally, genotypes from the GliomaScan GWAS can be accessed through dbGaP accession phs000652.v1.p1. Data from the other studies are available upon request.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Authors

Beatrice S Melin[1,41], Jill S Barnholtz-Sloan[2,41], Margaret R Wrensch[3,4,41], Christoffer Johansen[5,41], Dora Il'yasova[6,7,8,41], Ben Kinnersley[9,41], Quinn T Ostrom[2], Karim Labreche[9,10], Yanwen Chen[2], Georgina Armstrong[11], Yanhong Liu[11], Jeanette E Eckel-Passow[12], Paul A Decker[12], Marianne Labussière[10], Ahmed Idbaih[10,13], Khe Hoang-Xuan[10,13], Anna-Luisa Di Stefano[10,13], Karima Mokhtari[10,13], Jean-Yves Delattre[10,13], Peter Broderick[9], Pilar Galan[14], Konstantinos Gousias[15], Johannes Schramm[15], Minouk J Schoemaker[9], Sarah J Fleming[16], Stefan Herms[16], Stefanie Heilmann[17], Markus M Nöthen[17], Heinz-Erich Wichmann[18,19,20], Stefan Schreiber[21], Anthony Swerdlow[9,22], Mark Lathrop[23], Matthias Simon[15], Marc Sanson[10,13], Ulrika Andersson[1], Preetha Rajaraman[24], Stephen Chanock[24], Martha Linet[24], Zhaoming Wang[24], Meredith Yeager[24], GliomaScan Consortium[25], John K Wiencke[3,4], Helen Hansen[3], Lucie McCoy[3], Terri Rice[3], Matthew L Kosel[12], Hugues Sicotte[12], Christopher I Amos[26], Jonine L Bernstein[27], Faith Davis[28], Dan Lachance[29], Ching Lau[30], Ryan T Merrell[31], Joellen Shildkraut[7,8], Francis Ali-Osman[7,32], Siegal Sadetzki[33,34], Michael Scheurer[30], Sanjay Shete[35], Rose K Lai[36,42], Elizabeth B Claus[37,38,42], Sara H Olson[27,42], Robert B Jenkins[39,42], Richard S Houlston[9,40,42], and Melissa L Bondy[11,42]

Affiliations

[1]Department of Radiation Sciences, Umeå University, Umeå, Sweden

[2]Case Comprehensive Cancer Center, School of Medicine, Case Western Reserve University, Cleveland, Ohio, USA

[3]Department of Neurological Surgery, School of Medicine, University of California, San Francisco, San Francisco, California, USA

[4]Institute of Human Genetics, University of California, San Francisco, San Francisco, California, USA

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

[5]Institute of Cancer Epidemiology, Danish Cancer Society, Copenhagen, Denmark and Rigshospitalet, University of Copenhagen, Copenhagen, Denmark

[6]Department of Epidemiology and Biostatistics, School of Public Health, Georgia State University, Atlanta, Georgia, USA

[7]Duke Cancer Institute, Duke University Medical Center, Durham, North Carolina, USA

[8]Cancer Control and Prevention Program, Department of Community and Family Medicine, Duke University Medical Center, Durham, North Carolina, USA

[9]Division of Genetics and Epidemiology, Institute of Cancer Research, London, UK

[10]Sorbonne Universités UPMC Univ Paris 06, INSERM CNRS, U1127, UMR 7225, ICM, Paris, France

[11]Department of Medicine, Dan L. Duncan Comprehensive Cancer Center, Baylor College of Medicine, Houston, Texas, USA

[12]Division of Biomedical Statistics and Informatics, Mayo Clinic College of Medicine, Rochester, Minnesota, USA

[13]AP-HP, Groupe Hospitalier Pitié-Salpêtrière, Service de neurologie 2-Mazarin, Paris, France

[14]Université Paris 13 Sorbonne Paris Cité, INSERM U557, INRA U1125, CNAM, Paris, France

[15]Department of Neurosurgery, University of Bonn Medical Center, Bonn, Germany

[16]Centre for Epidemiology and Biostatistics, Faculty of Medicine and Health, University of Leeds, Leeds, UK

[17]Institute of Human Genetics, University of Bonn, Bonn, Germany

[18]Helmholtz Center Munich, Institute of Epidemiology I, Munich, Germany

[19]Institute of Medical Informatics, Biometry and Epidemiology, Ludwig Maximilians University, Munich, Germany

[20]Institute of Medical Statistics and Epidemiology, Technical University Munich, Munich, Germany

[21]1st Medical Department, University Clinic Schleswig–Holstein, Campus Kiel, Kiel, Germany

[22]Division of Breast Cancer Research, Institute of Cancer Research, London, UK

[23]Génome Québec, Department of Human Genetics, McGill University, Montreal, Quebec, Canada

[24]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, Maryland, USA

[26]Department of Biomedical Data Science, Geisel School of Medicine at Dartmouth, Hanover, New Hampshire, USA

[27]Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, New York, USA

[28]School of Public Health, University of Alberta, Edmonton, Alberta, Canada

[29]Department of Neurology, Mayo Clinic Comprehensive Cancer Center, Mayo Clinic, Rochester, Minnesota, USA

[30]Department of Pediatrics, Dan L. Duncan Comprehensive Cancer Center, Baylor College of Medicine, Houston, Texas, USA

[31]Department of Neurology, NorthShore University HealthSystem, Evanston, Illinois, USA

[32]Department of Surgery, Duke University Medical Center, Durham, North Carolina, USA

[33]Cancer and Radiation Epidemiology Unit, Gertner Institute, Chaim Sheba Medical Center, Tel Hashomer, Israel

[34]Department of Epidemiology and Preventive Medicine, School of Public Health, Sackler Faculty of Medicine, Tel-Aviv University, Tel-Aviv, Israel

[35]Department of Biostatistics, University of Texas Maryland Anderson Cancer Center, Houston, Texas, USA

[36]Departments of Neurology and Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, California, USA

[37]School of Public Health, Yale University, New Haven, Connecticut, USA

[38]Department of Neurosurgery, Brigham and Women’s Hospital, Boston, Massachusetts, USA

[39]Department of Laboratory Medicine and Pathology, Mayo Clinic Comprehensive Cancer Center, Mayo Clinic, Rochester, Minnesota, USA

[40]Division of Molecular Pathology, Institute of Cancer Research, London, UK

Acknowledgments

We are grateful to all of the patients and individuals for their participation, and we would also like to thank the clinicians and other hospital staff members, cancer registries and the study staff members in the respective centers who contributed to the blood sample and data collection.

The GICC was supported by grants from the US National Institutes of Health (NIH) (R01CA139020 (M.L.B. and B.S.M.), R01CA52689 (M.R.W.), R01CA52689 (M.L.B.) and P30CA125123 (M. Scheurer). Additional support was provided by the McNair Medical Institute (M. Scheurer) and the Population Sciences Biorepository at Baylor College of Medicine (M. Scheurer).

In Sweden, work was additionally supported by Acta Oncologica through the Royal Swedish Academy of Science (B.S.M.’s salary) and by the Swedish Research Council (B.S.M.) and the Swedish Cancer Foundation (B.S.M.). We are grateful to the National Clinical Brain Tumor Group and to all of the clinicians and research nurses throughout Sweden who identified all of the cases.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

In the UK, funding was provided by Cancer Research UK (C1298/A8362 supported by the Bobby Moore Fund (R.S.H., B.K. and P.B.), the Wellcome Trust (R.S.H., B.K. and P.B.) and the DJ Fielding Medical Research Trust (R.S.H., B.K. and P.B.). The National Brain Tumor Study is supported by the National Cancer Research Network, and we acknowledge all clinicians and healthcare professionals who contributed to this initiative. The UK INTERPHONE study was supported by the European Union Fifth Framework Program 'Quality of Life and Management of Living Resources' (QLK4-CT-1999-01563) (A.S., M.J.S. and S.J.F.) and the International Union against Cancer (UICC) (A.S., M.J.S. and S.J.F.). The UICC received funds from the Mobile Manufacturers' Forum and the GSM Association. Provision of funds via the UICC was governed by agreements that guaranteed INTERPHONE's scientific independence (http://www.iarc.fr/ENG/Units/RCAd.html), and the views expressed in the paper are not necessarily those of the funders. The UK centers were also supported by the Mobile Telecommunications and Health Research (MTHR) Programme, and the Northern UK Centre (A.S., M.J.S. and S.J.F.) was supported by the Health and Safety Executive, Department of Health and Safety Executive and the UK Network Operators.

In France, funding was provided by the Ligue Nationale Contre le Cancer (J.-Y.D.), the Fondation ARC (M. Sanson), the Institut National du Cancer (INCa; PL046; (M. Sanson)), the French Ministry of Higher Education and Research and the program "Investissements d'avenir" ANR-10-IAIHU-06 (M. Sanson, J.-Y.D., M. Labussière, A.-L.D.S., P.G., K.M., A.I., K.H.-X. and K.L.). This study was additionally supported by a grant from Génome Québec, le Ministère de l'Enseignement supérieur, de la Recherche, de la Science et de la Technologie (MESRST) Québec (M. Lathrop) and McGill University (M. Lathrop).

In Germany, funding was provided to M. Simon and J. Schramm by the Deutsche Forschungsgemeinschaft (Si552, Schr285), the Deutsche Krebshilfe (70-2385-Wi2, 70-3163-Wi3, 10-6262) and BONFOR. Funding for the WTCCC was provided by the Wellcome Trust (076113 and 085475; M. Simon and J. Schramm). The KORA Ausburg studies are supported by grants from the German Federal Ministry of Education and Research (BMBF) and were mainly financed by the Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg. This work was financed by the German National Genome Research Network (NGFN) (S. Schreiber and H.E.-W.) and supported within the Munich Center of Health Sciences (MC Health) as part of LMUinnovativ (S. Schreiber and H.E.-W.). Generation of the German control data was partially supported by a grant of the German Federal Ministry of Education and Research (BMBF) through the Integrated Network IntegraMent (Integrated Understanding of Causes and Mechanisms in Mental Disorders), under the auspices of the e:Med research and funding concept (01ZX1314A) (M.M.N., S. Herms and S. Heilmann). M.M.N. is a member of the DFG-funded Excellence Cluster ImmunoSensation and received support from the Alfried Krupp von Bohlen und Halbach-Stiftung.

For the UK GWAS, we acknowledge the funders, organizations and individuals who contributed to the blood sample and data collection as listed in Hepworth *et al.*[45]. MD Anderson acknowledges the work of P. Adatto, F. Morice, H. Zhang, V. Levin, A.W.K. Yung, M. Gilbert, R. Sawaya, V. Puduvalli, C. Conrad, F. Lang and J. Weinberg from the Brain and Spine Center for the MDA GWAS. For the French study, we are indebted to A. Rahimian (Onconeurotek), A.M. Lekieffre and M. Brandel for help in collecting data and to Y. Marie for database support. For the German study, we are indebted to B. Harzheim (Bonn), S. Ott and A. Müller-Erkwoh (Bonn) for help with the acquisition of clinical data and to R. Mahlberg (Bonn), who provided technical support. The UK study made use of control genotyping data generated by the Wellcome Trust Case–Control Consortium. A full list of the investigators who contributed to the generation of the data is available from http://www.wtccc.org.uk. The MDA GWAS made use of control genotypes from the CGEMS prostate and breast cancer studies. A full list of the investigators who contributed to the generation of the data is available from http://cgems.cancer.gov/. French controls were taken from the SU.VI.MAX study. The German GWAS made use of genotyping data from three population control sources: KORA-gen39, the Heinz-Nixdorf RECALL study and POPGEN. The HNR cohort was established with the support of the Heinz-Nixdorf Foundation. F.D. received support from the BONFOR Programme of the University of Bonn, Germany.

The UCSF Adult Glioma Study was supported by the NIH (grant numbers R01CA52689 (M.R.W. and J.K.W.), P50CA097257 (M.R.W. and J.K.W.), R01CA126831 (J.K.W.) and R01CA139020 (M.R.W.)), the Loglio Collective (M.R.W. and J.K.W.), the National Brain Tumor Foundation (M.R.W.), the Stanley D. Lewis and Virginia S. Lewis Endowed Chair in Brain Tumor Research (M.R.W.), the Robert Magnin Newman Endowed Chair in Neuro-oncology (J.K.W.) and by donations from the families and friends of J. Berardi, H. Glaser, E. Olsen, R.E. Cooper and W. Martinusen. This project also was supported by the National Center for Research Resources and the National Center for Advancing Translational Sciences, NIH, through UCSF–CTSI grant UL1 RR024131 (UCSF CTSI). The contents of this work are solely the responsibility of the authors and do not necessarily represent the official views of the NIH. The collection of cancer incidence data used in this study was supported by the California Department of Public Health as part of the statewide cancer reporting program mandated by California Health and Safety Code section 103885, the National Cancer Institute's Surveillance, Epidemiology and End Results Program under contract HHSN261201000140C (awarded to the Cancer Prevention Institute of California), contract HHSN261201000035C (awarded to the University of Southern California) and contract HHSN261201000034C (awarded to the Public Health Institute), and the Centers for Disease Control and Prevention's National Program of Cancer Registries under agreement # U58DP003862-01 (awarded to the California Department of Public Health).

The ideas and opinions expressed herein are those of the author(s), and endorsement by the State of California Department of Public Health, the National Cancer Institute and the Centers for Disease Control and Prevention, or their contractors and subcontractors, is not intended nor should be inferred. Other significant contributors for the UCSF Adult Glioma Study include M. Berger, P. Bracci, S. Chang, J. Clarke, A. Molinaro, A. Perry, M. Pezmecki, M. Prados, I. Smirnov, T. Tihan, K. Walsh, J. Wiemels and S. Zheng.

At Mayo, the authors wish to acknowledge the study participants and the clinicians and research staff at the participating medical centers, the Mayo Clinic Biobank and Biospecimens Accessioning and Processing Shared Resource (in particular its manager, M. Cicek). Work at the Mayo Clinic beyond the GICC was also supported by the NIH (grants P50CA108961 (B. O'Neill) and P30CA15083 (R. Diasio)), the National Institute of Neurological Disorders and Stroke (grant RC1NS068222Z (R.B.J.)), the Bernie and Edith Waterman Foundation (R.B.J.) and the Ting Tsung and Wei Fong Chao Family Foundation (R.B.J.).

The GliomaScan Consortium comprised (apart from authors listed in the author list): L.E.B. Freeman (Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, Maryland, USA), S. Koutros (Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, Maryland, USA), D. Albanes (Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, Maryland, USA), K. Visvanathan (Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA and Johns Hopkins Sidney Kimmel Comprehensive Cancer Center, Baltimore, Maryland, USA), V.L. Stevens (Epidemiology Research Program, American Cancer Society, Atlanta, Georgia, USA), R. Henriksson (Department of Radiation Sciences, Oncology, Umea University, Umea, Sweden), D.S. Michaud (Department of Public Health and Community Medicine, Tufts University Medical School, Boston, Massachusetts, USA), M. Feychting (Unit of Epidemiology, Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden), A. Ahlbom (Unit of Epidemiology, Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden), G.G. Giles (Cancer Epidemiology Centre, Cancer Council of Victoria, Melbourne, Victoria, Australia and Centre for Molecular, Environmental, Genetic and Analytic Epidemiology, University of Melbourne, Melbourne, Victoria, Australia), R. Milne (Cancer Epidemiology Centre, Cancer Council of Victoria, Melbourne, Victoria, Australia and Centre for Molecular, Environmental, Genetic and Analytic Epidemiology, University of Melbourne, Melbourne, Victoria, Australia), R. McKean-Cowdin (Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, California, USA), L. Le Marchand (Cancer Research Center, University of Hawaii, Honolulu, Hawaii, USA), M. Stampfer (Channing Laboratory, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA and Departments of Epidemiology and Nutrition, Harvard School of Public Health, Boston, Massachusetts, USA), A.M. Ruder (National Institute for Occupational Safety and Health, Centers for Disease Control and Prevention, Cincinnati, Ohio, USA), T. Carreon (National Institute for Occupational Safety and Health, Centers for Disease Control and Prevention, Cincinnati, Ohio, USA), G. Hallmans (Department of Public Health and Clinical Medicine/Nutritional Research, Umea University, Umea, Sweden), A. Zeleniuch-Jacquotte (Division of Epidemiology, Department of Environmental Medicine, New York University School of Medicine, New York, New York, USA), J.M. Gaziano (Massachusetts Veteran's Epidemiology, Research and Information Center, Geriatric Research Education and Clinical Center, VA Boston Healthcare System, Boston, Massachusetts, USA), H.D. Sesso (Division of Preventive Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA), M.P. Purdue (Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, Maryland, USA), E. White (Fred Hutchinson Cancer Research Center, Seattle, Washington, USA and Department of Epidemiology, University of Washington, Seattle, Washington, USA) and J. Buring (Division of Preventive Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA).

UK10K data generation and access was organized by the UK10K consortium and funded by the Wellcome Trust.

References

1. Bondy ML, et al. Brain tumor epidemiology: consensus from the Brain Tumor Epidemiology Consortium. Cancer. 2008; 113:1953–1968. [PubMed: 18798534]
2. Ostrom QT, et al. CBTRUS statistical report: primary brain and central nervous system tumors diagnosed in the United States in 2008–2012. Neuro-oncol. 2015; 17(Suppl. 4):iv1–iv62. [PubMed: 26511214]
3. Louis DN, et al. The 2007 WHO classification of tumors of the central nervous system. Acta Neuropathol. 2007; 114:97–109. [PubMed: 17618441]
4. Ostrom QT, et al. CBTRUS statistical report: primary brain and central nervous system tumors diagnosed in the United States in 2007–2011. Neuro-oncol. 2014; 16:iv1–iv63. [PubMed: 25304271]
5. Ostrom QT, et al. The epidemiology of glioma in adults: a 'state of the science' review. Neuro-oncol. 2014; 16:896–913. [PubMed: 24842956]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

6. Hemminki K, Tretli S, Sundquist J, Johannesen TB, Granstrom C. Familial risks in nervous-system tumors: a histology-specific analysis from Sweden and Norway. Lancet Oncol. 2009; 10:481–488. [PubMed: 19356978]
7. Sanson M, et al. Chromosome 7p11.2 (*EGFR*) variation influences glioma risk. Hum. Mol. Genet. 2011; 20:2897–2904. [PubMed: 21531791]
8. Shete S, et al. Genome-wide association study identifies five susceptibility loci for glioma. Nat. Genet. 2009; 41:899–904. [PubMed: 19578367]
9. Wrensch M, et al. Variants in the *CDKN2B* and *RTEL1* regions are associated with high-grade glioma susceptibility. Nat. Genet. 2009; 41:905–908. [PubMed: 19578366]
10. Kinnersley B, et al. Genome-wide association study identifies multiple susceptibility loci for glioma. Nat. Commun. 2015; 6:8559. [PubMed: 26424050]
11. Walsh KM, et al. Variants near *TERT* and *TERC* influencing telomere length are associated with high-grade glioma risk. Nat. Genet. 2014; 46:731–735. [PubMed: 24908248]
12. Jenkins RB, et al. A low-frequency variant at 8q24.21 is strongly associated with risk of oligodendroglial tumors and astrocytomas with *IDH1* or *IDH2* mutation. Nat. Genet. 2012; 44:1122–1125. [PubMed: 22922872]
13. Rajaraman P, et al. Genome-wide association study of glioma and meta-analysis. Hum. Genet. 2012; 131:1877–1888. [PubMed: 22886559]
14. Stacey SN, et al. A germline variant in the *TP53* polyadenylation signal confers cancer susceptibility. Nat. Genet. 2011; 43:1098–1103. [PubMed: 21946351]
15. Kinnersley B, et al. Quantifying the heritability of glioma using genome-wide complex trait analysis. Sci. Rep. 2015; 5:17267. [PubMed: 26625949]
16. Lonsdale J, et al. The Genotype–Tissue Expression (GTEx) project. Nat. Genet. 2013; 45:580–585. [PubMed: 23715323]
17. Westra HJ, et al. Systematic identification of *trans* eQTLs as putative drivers of known disease associations. Nat. Genet. 2013; 45:1238–1243. [PubMed: 24013639]
18. Zhu Z, et al. Integration of summary data from GWAS and eQTL studies predicts complex-trait gene targets. Nat. Genet. 2016; 48:481–487. [PubMed: 27019110]
19. Andersson R, et al. An atlas of active enhancers across human cell types and tissues. Nature. 2014; 507:455–461. [PubMed: 24670763]
20. Hnisz D, et al. Super-enhancers in the control of cell identity and disease. Cell. 2013; 155:934–947. [PubMed: 24119843]
21. Ruark E, et al. Identification of nine new susceptibility loci for testicular cancer, including variants near *DAZL* and *PRDM14*. Nat. Genet. 2013; 45:686–689. [PubMed: 23666240]
22. Wu C, et al. Genome-wide association analyses of esophageal squamous cell carcinoma in Chinese identify multiple susceptibility loci and gene–environment interactions. Nat. Genet. 2012; 44:1090–1097. [PubMed: 22960999]
23. Riemenschneider MJ, et al. Amplification and overexpression of the *MDM4* (*MDMX*) gene from 1q32 in a subset of malignant gliomas without *TP53* mutation or *MDM2* amplification. Cancer Res. 1999; 59:6091–6096. [PubMed: 10626796]
24. Boland E, et al. Mapping of deletion and translocation breakpoints in 1q44 implicates the serine–threonine kinase AKT3 in postnatal microcephaly and agenesis of the corpus callosum. Am. J. Hum. Genet. 2007; 81:292–303. [PubMed: 17668379]
25. Turner KM, et al. Genomically amplified Akt3 activates DNA repair pathway and promotes glioma progression. Proc. Natl. Acad. Sci. USA. 2015; 112:3421–3426. [PubMed: 25737557]
26. Gur G, et al. LRIG1 restricts growth factor signaling by enhancing receptor ubiquitylation and degradation. EMBO J. 2004; 23:3270–3281. [PubMed: 15282549]
27. Yang JA, et al. LRIG1 enhances the radio-sensitivity of radio-resistant human glioblastoma U251 cells via attenuation of the EGFR–AKT signaling pathway. Int. J. Clin. Exp. Pathol. 2015; 8:3580–3590. [PubMed: 26097540]
28. Wei J, et al. miR-20a mediates temozolomide resistance in glioblastoma cells via negatively regulating LRIG1 expression. Biomed. Pharmacother. 2015; 71:112–118. [PubMed: 25960225]

29. Watanabe T, Nobusawa S, Kleihues P, Ohgaki H. *IDH1* mutations are early events in the development of astrocytomas and oligodendrogliomas. Am. J. Pathol. 2009; 174:1149–1153. [PubMed: 19246647]

30. Yan H, et al. *IDH1* and *IDH2* mutations in gliomas. N. Engl. J. Med. 2009; 360:765–773. [PubMed: 19228619]

31. Noushmehr H, et al. Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. Cancer Cell. 2010; 17:510–522. [PubMed: 20399149]

32. Christensen BC, et al. DNA methylation, isocitrate dehydrogenase mutation and survival in glioma. J. Natl. Cancer Inst. 2011; 103:143–153. [PubMed: 21163902]

33. Sanson M, et al. Isocitrate dehydrogenase 1 codon 132 mutation is an important prognostic biomarker in gliomas. J. Clin. Oncol. 2009; 27:4150–4154. [PubMed: 19636000]

34. Eckel-Passow JE, et al. Glioma groups based on 1p/19q, *IDH* and *TERT* promoter mutations in tumors. N. Engl. J. Med. 2015; 372:2499–2508. [PubMed: 26061753]

35. Walsh KM, et al. Telomere maintenance and the etiology of adult glioma. Neuro-oncol. 2015; 17:1445–1452. [PubMed: 26014050]

36. Miyake Y, et al. RPA-like mammalian Ctc1–Stn1–Ten1 complex binds to single-stranded DNA and protects telomeres independently of the Pot1 pathway. Mol. Cell. 2009; 36:193–206. [PubMed: 19854130]

37. Chen LY, Redon S, Lingner J. The human CST complex is a terminator of telomerase activity. Nature. 2012; 488:540–544. [PubMed: 22763445]

38. Bainbridge MN, et al. Germline mutations in shelterin complex genes are associated with familial glioma. J. Natl. Cancer Inst. 2014; 107:384. [PubMed: 25482530]

39. Zhang C, et al. Genetic determinants of telomere length and risk of common cancers: a mendelian randomization study. Hum. Mol. Genet. 2015; 24:5356–5366. [PubMed: 26138067]

40. Walsh KM, et al. Longer genotypically estimated leukocyte telomere length is associated with increased adult glioma risk. Oncotarget. 2015; 6:42468–42477. [PubMed: 26646793]

41. Ceccarelli M, et al. Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma. Cell. 2016; 164:550–563. [PubMed: 26824661]

42. Xipell E, et al. Endoplasmic-reticulum-stress-inducing drugs sensitize glioma cells to temozolomide through downregulation of MGMT, MPG and Rad51. Neuro-oncol. 2016; 18:1109–1119. [PubMed: 26951384]

43. Nicolas CS, et al. The role of JAK–STAT signaling within the CNS. JAK-STAT. 2013; 2:e22925. [PubMed: 24058789]

44. Cancer Genome Atlas Research Network. Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. N. Engl. J. Med. 2015; 372:2481–2498. [PubMed: 26061751]

45. Hepworth SJ, et al. Mobile phone use and risk of glioma in adults: case–control study. BMJ. 2006; 332:883–887. [PubMed: 16428250]

46. Amirian ES, et al. The Glioma International Case–Control Study: a report from the Genetic Epidemiology of Glioma International Consortium. Am. J. Epidemiol. 2016; 183:85–91. [PubMed: 26656478]

47. Li Y, et al. FastPop: a rapid principal component–derived method to infer intercontinental ancestry using genetic data. BMC Bioinformatics. 2016; 17:122. [PubMed: 26961892]

48. Cardis E, et al. The INTERPHONE study: design, epidemiological methods and description of the study population. Eur. J. Epidemiol. 2007; 22:647–664. [PubMed: 17636416]

49. Hercberg S, et al. The SU.VI.MAX study: a randomized, placebo-controlled trial of the health effects of antioxidant vitamins and minerals. Arch. Intern. Med. 2004; 164:2335–2342. [PubMed: 15557412]

50. Wichmann HE, Gieger C, Illig T. MONICA–KORA Study Group. KORA-gen—resource for population genetics, controls and a broad spectrum of disease phenotypes. Gesundheitswesen. 2005; 67(Suppl. 1):S26–S30. [PubMed: 16032514]

51. Krawczak M, et al. PopGen: population-based recruitment of patients and controls for the analysis of complex genotype–phenotype relationships. Community Genet. 2006; 9:55–61. [PubMed: 16490960]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

52. Schmermund A, et al. Assessment of clinically silent atherosclerotic disease, and established and novel risk factors for predicting myocardial infarction and cardiac death in healthy middle-aged subjects: rationale and design of the Heinz Nixdorf RECALL study. Risk factors, evaluation of coronary calcium and lifestyle. Am. Heart J. 2002; 144:212–218. [PubMed: 12177636]
53. Hunter DJ, et al. A genome-wide association study identifies alleles in *FGFR2* associated with risk of sporadic postmenopausal breast cancer. Nat. Genet. 2007; 39:870–874. [PubMed: 17529973]
54. Yeager M, et al. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. Nat. Genet. 2007; 39:645–649. [PubMed: 17401363]
55. Wiemels JL, et al. History of allergies among adults with glioma and controls. Int. J. Cancer. 2002; 98:609–615. [PubMed: 11920623]
56. Felini MJ, et al. Reproductive factors and hormone use risk of adult gliomas. Cancer Causes Control. 2009; 20:87–96. [PubMed: 18766447]
57. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 2009; 5:e1000529. [PubMed: 19543373]
58. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. Nature. 2010; 467:1061–1073. [PubMed: 20981092]
59. Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat. Genet. 2007; 39:906–913. [PubMed: 17572673]
60. Purcell S, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 2007; 81:559–575. [PubMed: 17701901]
61. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet. 2006; 2:e190. [PubMed: 17194218]
62. Jenkins RB, et al. Distinct germline polymorphisms underlie glioma morphologic heterogeneity. Cancer Genet. 2011; 204:13–18. [PubMed: 21356187]
63. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000; 155:945–959. [PubMed: 10835412]
64. Liu JZ, et al. Meta-analysis and imputation refines the association of 15q25 with smoking quantity. Nat. Genet. 2010; 42:436–440. [PubMed: 20418889]
65. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. Br. Med. J. 2003; 327:557–560. [PubMed: 12958120]
66. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex-trait analysis. Am. J. Hum. Genet. 2011; 88:76–82. [PubMed: 21167468]
67. Yang J, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nat. Genet. 2012; 44:369–375. [PubMed: 22426310]
68. Ward LD, Kellis M. HaploReg: a resource for exploring chromatin states, conservation and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Res. 2012; 40:D930–D934. [PubMed: 22064851]
69. Boyle AP, et al. Annotation of functional variation in personal genomes using RegulomeDB. Genome Res. 2012; 22:1790–1797. [PubMed: 22955989]
70. Ng SB, et al. Targeted capture and massively parallel sequencing of 12 human exomes. Nature. 2009; 461:272–276. [PubMed: 19684571]
71. Roadmap Epigenomics Consortium. Integrative analysis of 111 reference human epigenomes. Nature. 2015; 518:317–330. [PubMed: 25693563]
72. Pharoah PD, Antoniou AC, Easton DF, Ponder BA. Polygenes, risk prediction and targeted prevention of breast cancer. N. Engl. J. Med. 2008; 358:2796–2803. [PubMed: 18579814]
73. Scheurer ME, et al. Familial aggregation of glioma: a pooled analysis. Am. J. Epidemiol. 2010; 172:1099–1107. [PubMed: 20858744]
74. Danecek P, et al. The variant call format and VCFtools. Bioinformatics. 2011; 27:2156–2158. [PubMed: 21653522]

75. Scales M, Jager R, Migliorini G, Houlston RS, Henrion MY. visPIG—a web tool for producing multiregion, multitrack, multiscale plots of genetic data. PLoS One. 2014; 9:e107497. [PubMed: 25208325]

76. Gabriel SB, et al. The structure of haplotype blocks in the human genome. Science. 2002; 296:2225–2229. [PubMed: 12029063]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

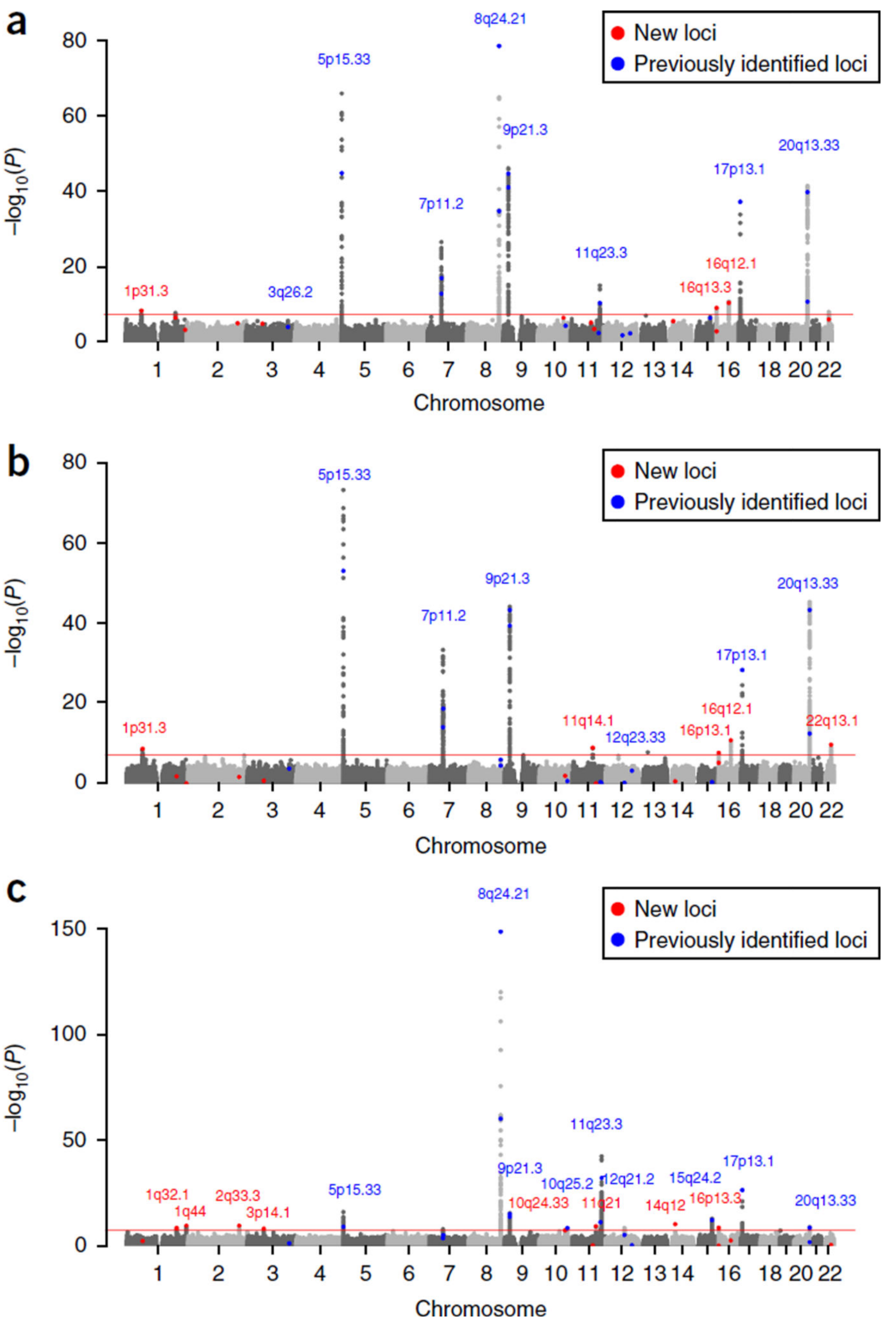


Figure 1.

Genome-wide discovery-phase meta-analysis P-values ($-\log_{10}P$) plotted against their chromosomal positions. (**a**) All glioma. (**b**) GBM. (**c**) Non-GBM tumors. The red horizontal line corresponds to a significance threshold of $P = 5.0 \times 10^{-8}$. New and known loci are labeled in red and blue, respectively.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

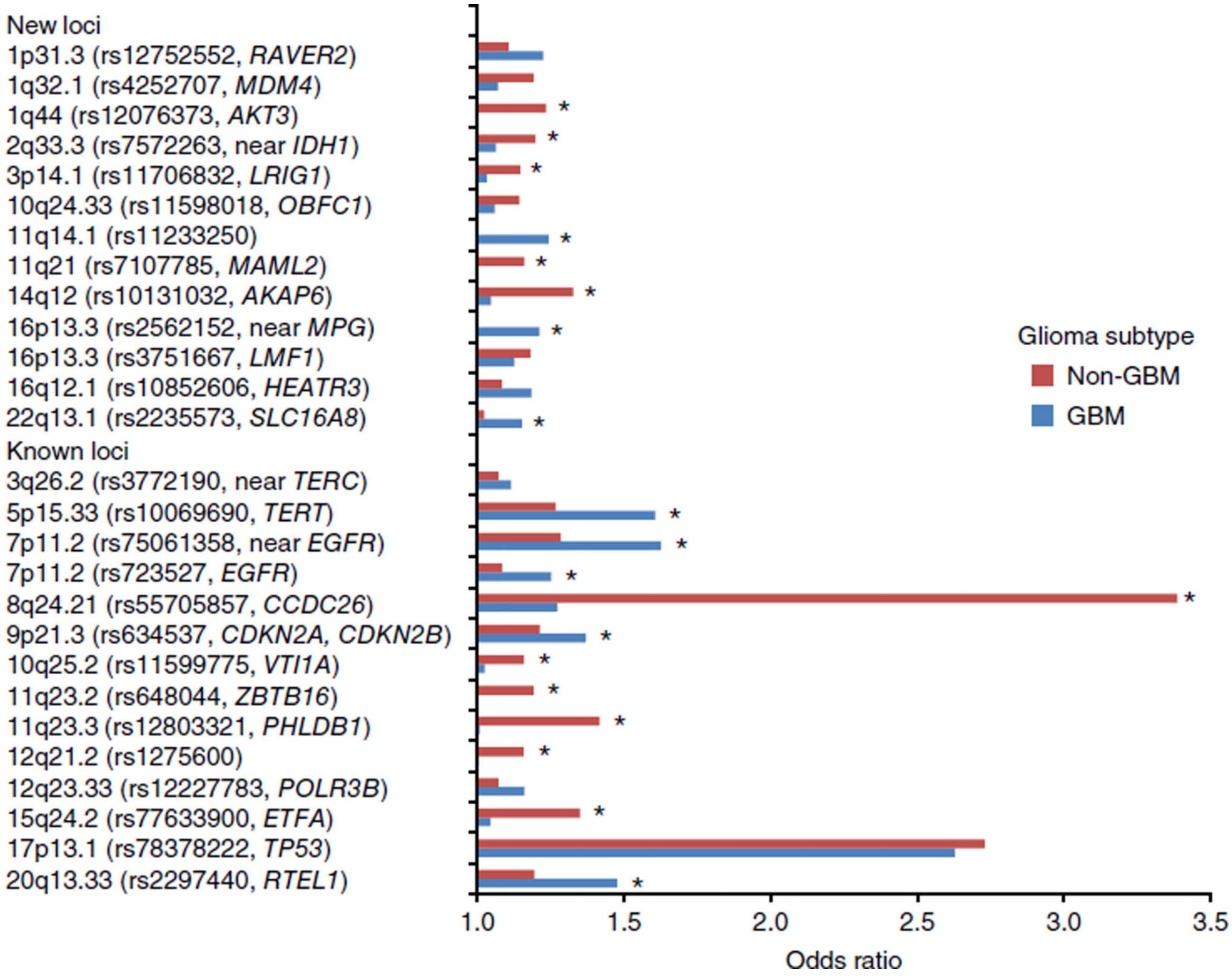


Figure 2.

Relative impact of SNP associations at known and newly identified risk loci for GBM and non-GBM tumors. Odds ratios (ORs) derived with respect to the risk allele. Asterisks denote SNPs showing a significant difference between GBM and non-GBM tumors from the case-only analysis as detailed in Supplementary Table 4.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 1

Association statistics for the top SNP at each of the newly-reported glioma risk loci

Locus	Subtype	SNP	Position	Alleles	RAF	INFO	All glioma *P*	All glioma OR (95% CI)	GBM glioma *P*	GBM glioma OR (95% CI)	Non-GBM glioma *P*	Non-GBM glioma OR (95% CI)
1p31.3	GBM	rs12752552	65229299	**T**/C	0.870	0.992	**4.07×10^{-9}**	1.18 (1.11–1.24)	**2.04×10^{-9}**	1.22 (1.15–1.31)	4.78×10^{-3}	1.11 (1.03–1.18)
1q32.1	Non-GBM	rs4252707	204508147	G/**A**	0.220	0.992	2.97×10^{-7}	1.12 (1.07–1.17)	0.015	1.07 (1.01–1.13)	**3.34×10^{-9}**	1.19 (1.12–1.26)
1q44	Non-GBM	rs12076373	243851947	**G**/C	0.837	0.996	4.97×10^{-4}	1.09 (1.04–1.15)	0.846	0.99 (0.94–1.06)	**2.63×10^{-10}**	1.23 (1.16–1.32)
2q33.3	Non-GBM	rs7572263	209051586	**A**/G	0.756	0.997	2.58×10^{-6}	1.11 (1.06–1.15)	0.019	1.06 (1.01–1.12)	**2.18×10^{-10}**	1.20 (1.13–1.26)
3p14.1	Non-GBM	rs11706832	66502981	A/**C**	0.456	0.997	1.06×10^{-5}	1.08 (1.05–1.12)	0.158	1.03 (0.99–1.08)	**7.66×10^{-9}**	1.15 (1.09–1.20)
10q24.33	Non-GBM	rs11598018	105661315	**C**/A	0.462	0.960	3.07×10^{-7}	1.10 (1.06–1.14)	0.0103	1.06 (1.01–1.11)	**3.39×10^{-8}**	1.14 (1.09–1.20)
11q14.1	GBM	rs11233250	82397014	**C**/T	0.868	0.990	5.40×10^{-6}	1.14 (1.08–1.21)	**9.95×10^{-10}**	1.24 (1.16–1.33)	0.592	0.98 (0.91–1.05)
11q21	Non-GBM	rs7107785	95747337	**T**/C	0.479	0.997	2.96×10^{-4}	1.07 (1.03–1.11)	0.844	1.00 (0.95–1.04)	**3.87×10^{-10}**	1.16 (1.11–1.21)
14q12	Non-GBM	rs10131032	33250081	**G**/A	0.916	0.991	2.33×10^{-6}	1.17 (1.09–1.24)	0.247	1.05 (0.97–1.13)	**5.07×10^{-11}**	1.33 (1.22–1.44)
16p13.3	GBM	rs2562152	123896	A/**T**	0.850	0.937	1.18×10^{-3}	1.09 (1.04–1.15)	**1.93×10^{-8}**	1.21 (1.13–1.29)	0.948	1.00 (0.93–1.07)
16p13.3	Non-GBM	rs3751667	1004554	C/**T**	0.208	0.985	**8.75×10^{-10}**	1.14 (1.09–1.19)	5.95×10^{-6}	1.13 (1.07–1.19)	**2.61×10^{-9}**	1.18 (1.12–1.25)
16q12.1	GBM	rs10852606	50128872	T/**C**	0.713	0.990	**3.66×10^{-11}**	1.14 (1.10–1.19)	**1.29×10^{-11}**	1.18 (1.13–1.24)	2.42×10^{-3}	1.08 (1.03–1.14)
22q13.1	GBM	rs2235573	38477930	**G**/A	0.507	0.995	8.64×10^{-7}	1.09 (1.06–1.13)	**1.76×10^{-10}**	1.15 (1.10–1.20)	0.325	1.02 (0.97–1.07)

Associations at $P < 5 \times 10^{-8}$ are highlighted in bold. Odds ratios (ORs) were derived with respect to the risk allele underlined and highlighted in bold. Risk allele frequency (RAF) is according to European samples from the 1000 Genomes project. The INFO column indicates the average imputation info score across all studies, with a score of 1 indicating that the SNP is directly genotyped in all studies. CI, confidence interval.

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/263582938

Chlamydia psittaci: New insights into genomic diversity, clinical pathology, host-pathogen interaction and anti-bacterial immunity

Article *in* International Journal of Medical Microbiology · October 2014

DOI: 10.1016/j.ijmm.2014.06.010

CITATIONS
85

READS
2,029

16 authors, including:

Angela Berndt
Friedrich-Loeffler-Institute, Gemany, Jena
88 PUBLICATIONS **2,255** CITATIONS
SEE PROFILE

Selina Böcker
Leibniz Institute for Natural Product Research and Infection Biology - Hans Knöll I…
3 PUBLICATIONS **111** CITATIONS
SEE PROFILE

Frank Hänel
Leibniz Institute for Natural Product Research and Infection Biology - Hans Knöll I…
57 PUBLICATIONS **2,309** CITATIONS
SEE PROFILE

All content following this page was uploaded by Hans-Peter Saluz on 30 July 2014.

The user has requested enhancement of the downloaded file.

International Journal of Medical Microbiology xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

International Journal of Medical Microbiology

journal homepage: www.elsevier.com/locate/ijmm

Mini Review

Chlamydia psittaci: New insights into genomic diversity, clinical pathology, host–pathogen interaction and anti-bacterial immunity

Q1 Michael R. Knittler[a,*], Angela Berndt[b], Selina Böcker[c], Pavel Dutow[d], Frank Hänel[c], Dagmar Heuer[e], Danny Kägebein[a], Andreas Klos[d], Sophia Koch[e], Elisabeth Liebler-Tenorio[b], Carola Ostermann[b], Petra Reinhold[b], Hans Peter Saluz[c], Gerhard Schöfl[c], Philipp Sehnert[c], Konrad Sachse[b,*]

[a] Institute of Immunology, Friedrich-Loeffler-Institut, Isle of Riems, Germany
[b] Institute of Molecular Pathogenesis, Friedrich-Loeffler-Institut, Jena, Germany
[c] Hans Knöll Institute, Leibniz Institute for Natural Product Research and Infection Biology, Jena, Germany
[d] Institute for Medical Microbiology and Hospital Epidemiology, Hannover Medical School, Hannover, Germany
[e] Robert Koch Institute, Berlin, Germany

ARTICLE INFO

Keywords:
Chlamydia psittaci
Animal models
Antigen presentation
Host adaptation
Immune response
Molecular pathogenesis

ABSTRACT

The distinctive and unique features of the avian and mammalian zoonotic pathogen *Chlamydia* (*C.*) *psittaci* include the fulminant course of clinical disease, the remarkably wide host range and the high proportion of latent infections that are not leading to overt disease. Current knowledge on associated diseases is rather poor, even in comparison to other chlamydial agents. In the present paper, we explain and summarize the major findings of a national research network that focused on the elucidation of host–pathogen interactions in vitro and in animal models of *C. psittaci* infection, with the objective of improving our understanding of genomics, pathology, pathophysiology, molecular pathogenesis and immunology and conceiving new approaches to therapy. We discuss new findings on comparative genome analysis, the complexity of pathophysiological interactions and systemic consequences, local immune response, the role of the complement system and antigen presentation pathways in the general context of state-of-the-art knowledge on chlamydial infections in humans and animals and single out relevant research topics to fill remaining knowledge gaps on this important yet somewhat neglected pathogen.

© 2014 Published by Elsevier GmbH.

UNCORRECTED PROOF

Introduction

The first definition of *Chlamydia* (*C.*) *psittaci* was formulated in 1968 to distinguish it from *C. trachomatis*, the only other known chlamydial species at that time, and was based on the inability to accumulate glycogen in its inclusions and resistance to sulfadiazine (Page, 1968). More recently, the species was redefined to comprise mainly avian strains, but at the same time assigned to the separate genus *Chlamydophila* (Everett et al., 1999). The latter has been abolished in the meantime (Kuo et al., 2011; Stephens et al., 2009), so that *C. psittaci* is now one of nine accepted members of the recombined genus *Chlamydia*. Two more avian chlamydiae (Sachse et al., 2014) and one Candidatus taxon (Vorimore et al., 2013) were discovered recently and are about to be added to the same genus.

The groundwork for modern chlamydia research was laid in the 1960s and 1970s by the pioneering work of J.W. Moulder, A. Matsumoto, G.P. Manire, T.P. Hatch, P.B. Wyrick and several other scientists, who focused on *C. psittaci* in their studies. By analyzing the structure and chemical composition of isolated *C. psittaci* "particles" (Moulder, 1962) conducted one of the first molecular characterizations of chlamydiae. Based on experiments with *C. psittaci*-infected cells (Matsumoto and Manire, 1970) were the first to provide high-resolution images of chlamydial bodies and to describe effects of antibiotics on their morphology. The groundbreaking studies of Hatch (1975) demonstrated that chlamydiae owe their obligate intracellular mode of reproduction to the requirement for energy intermediates produced by the host cell. Not least, Wyrick and colleagues (1978) were the first to describe structural properties of chlamydial compartments (Narita et al., 1976) and the ability of *C. psittaci* to infect immune cells (Wyrick and Brownridge, 1978).

C. psittaci is the causative agent of psittacosis (sometimes also called ornithosis or parrot fever), the most important animal

---

* Corresponding authors.
E-mail addresses: michael.knittler@fli.bund.de (M.R. Knittler), konrad.sachse@fli.bund.de (K. Sachse).

http://dx.doi.org/10.1016/j.ijmm.2014.06.010
1438-4221/© 2014 Published by Elsevier GmbH.

Please cite this article in press as: Knittler, M.R., et al., *Chlamydia psittaci*: New insights into genomic diversity, clinical pathology, host–pathogen interaction and anti-bacterial immunity. Int. J. Med. Microbiol. (2014), http://dx.doi.org/10.1016/j.ijmm.2014.06.010

chlamydiosis of zoonotic character. This systemic disease can take acute, protracted, chronic or subclinical courses and mainly affects psittacine birds and domestic poultry (Andersen and Vanrompay, 2000; Kaleta and Taday, 2003). Typical symptoms include lethargy, hyperthermia, abnormal excretions, respiratory distress, nasal and ocular discharges, weight loss and reduced egg production. *C. psittaci* infections and ensuing disease outbreaks are economically relevant to the poultry industry and significantly compromise animal welfare (Yin et al., 2013). Recent data also indicate that, in concert with *C. trachomatis*, *C. psittaci* may play a role in human trachoma (Dean et al., 2013).

The zoonotic dimension of avian chlamydiosis is another cause for concern, as cases of bird-to-man transmission are regularly reported in the literature (Gaede et al., 2008; Harkinezhad et al., 2009; Heddema et al., 2006; Laroucau et al., 2009). The fact that the overall number of reported cases remained generally low in the past decades is, to some extent, due to the absence of this pathogen in most routine diagnostic schemes. The symptoms in affected individuals are mainly non-specific and influenza-like, but severe pneumonia, endocarditis and encephalitis are not uncommon (Salisch et al., 1996). While it is generally perceived that the course and severity of disease depend on virulence properties of the *C. psittaci* strain involved, the immune response of the host, as well as transmission route and infectious dose, very little is known in concrete terms about any parameter of host–pathogen interactions. Furthermore, latent infection seems to be more widespread than clinical cases, both among birds and humans. It was shown in several studies in cattle that, in the absence of clinical signs, performance and health parameters of chlamydia carriers were clearly inferior to those of chlamydia-free animals (reviewed in Reinhold et al., 2011b). In this context, frequently observed recurrent infections by chlamydiae might pave the way for chronic disease.

The remarkable variety of clinical manifestations displayed by *C. psittaci* and chlamydial infections in general should, at least in part, be explicable through the distinctive features of the causative agent. *C. psittaci* is an obligately intracellular bacterium with a unique biphasic developmental cycle. It exists in basically two morphological stages, the infectious, but metabolically inactive elementary body (EB) and the non-infectious, but metabolically active reticulate body (RB). In the course of a replication cycle, EBs evolve into RBs in a vacuole-like inclusion of the host cell, where the latter undergo binary fission before transforming back into EBs to start a fresh cycle. The natural development can be halted at the intracellular stage through the action of adverse factors, such as interferon (IFN)-γ exposure, iron or amino acid depletion, and antibiotic treatment (Beatty et al., 1994; Wyrick, 2010). This persistent state is characterized by aberrant (enlarged) morphology of RBs, loss of cultivability and rescue of infectivity upon removal of inducers. In the case of *C. psittaci*, induction of persistence in vitro was accompanied by consistent downregulation of membrane proteins, chlamydial sigma factors, cell division protein and RB–EB differentiation proteins from 24 h post-infection onwards (Goellner et al., 2006). While it is not yet clear whether the host immune response can actually trigger transition of chlamydial bodies into the persistent state, although some evidence seems to suggest it (Borel et al., 2008; Pospischil et al., 2009), these observations provide an explanation for the failure of antibiotic treatment to eliminate chlamydial agents from an infected organism (Reinhold et al., 2011a).

Host preference of *C. psittaci* is an interesting yet largely unexplored topic. As the agent is generally referred to as an avian pathogen, there is still a widely held notion that it occurs only in birds. However, a number of surveys showed that *C. psittaci* can be found frequently in non-avian domestic animals, such as cattle (Kemmerling et al., 2009), sheep (Lenzko et al., 2011), swine (Kauffold et al., 2006; Vanrompay et al., 2004), horses (Szeredi et al., 2005; Theegarten et al., 2008), goats and cats (Pantchev et al., 2010), as well as in wildlife (Hotzel et al., 2004) and laboratory rodents (Henning et al., 2008). Although the role of the agent in these hosts remains to be clarified it seems certain that overt clinical manifestations are rare exceptions. In addition, the virulence to humans of non-avian *C. psittaci* strains appears to be negligible compared to avian isolates, since the reported cases of human psittacosis are usually traced back to contact with an avian source. Nevertheless, no genetic markers to distinguish between avian and non-avian strains have been found in comparative genomic studies (Voigt et al., 2012), nor have any virulence criteria been identified (Read et al., 2013). One can hypothesize from these findings that birds are the original and typical host for *C. psittaci*, while mammals merely act as alternate hosts and any passages of avian strains through non-avian hosts are accompanied by loss of virulence. Apart from these considerations, the factors determining host preference and adaptation of chlamydiae are largely unclear.

Comparative genomics: *C. psittaci* – a typical member of the family Chlamydiaceae with unique features

Second-generation sequencing efforts have led to the completion of more than 100 chlamydial genome sequences encompassing all recognized species, except *C. suis*. For *C. psittaci*, whole-genome sequences of 28 strains have become available so far (http://ftp.ncbi.nlm.nih.gov/genomes/ASSEMBLY_BACTERIA/Chlamydia_psittaci/), covering virtually all serotypes and avian and mammalian hosts (Grinblat-Huse et al., 2011; Schofl et al., 2011; Seth-Smith et al., 2011, 2013; Van Lent et al., 2012; Voigt et al., 2011). This wealth of data allows unprecedented insights in the genomic organization, evolutionary dynamics, and genetic diversity of the species.

Comparative analysis of chlamydial genomes showed a high level of sequence and gene order conservation (shared synteny) across members of the family and, as in other obligate intracellular bacteria, a considerably reduced genome, implying dependency on the host organism for many metabolic capabilities. All sequenced *C. psittaci* strains possess a single chromosome of approximately 1.1 Mb, and a conserved plasmid of ~8 kb containing 7–8 protein-coding sequences (CDS) was reported for 12 strains. A total of 911 core CDS are shared among all currently sequenced *C. psittaci* strains, which is equivalent to about 90% of the genes present in each of these genomes (Read et al., 2013). Among the species of *C. psittaci*, *C. abortus*, *C. pneumoniae*, and *C. trachomatis*, 736 genes are still shared (with a total CDS count ranging from 874 to 1097), which illustrates the degree of conservation across chlamydial genomes (Voigt et al., 2012). In contrast, one of the closest relative of *Chlamydiaceae* for which a genomic sequence is available, *Simkania negevensis*, features a genome of more than twice this size with a chromosome of 2.5 Mb and ~2500 predicted CDSs (Collingro et al., 2011).

Conserved synteny is extremely high among most chlamydial genomes. Major genomic rearrangements are present only between the closely related *C. trachomatis* and *C. muridarum* versus the remainder of the *Chlamydiaceae* (Fig. 1). Major deviations from synteny and/or sequence conservation, however, can be observed in genes encoding inclusion membrane proteins, in the hypervariable region near the predicted replication termination region known as the plasticity zone (PZ), as well as among the family of polymorphic membrane protein (*pmp*) genes (Fig. 1). These regions are assumed to harbor key genomic correlates to species-specific adaptation (see also Section "Host adaptation of *C. psittaci*: Escape from the avian innate immune response") to environmental niches, tissue tropism and differences in virulence and pathogenicity (Thomson et al., 2005).

UNCORRECTED PROOF

Please cite this article in press as: Knittler, M.R., et al., *Chlamydia psittaci*: New insights into genomic diversity, clinical pathology, host–pathogen interaction and anti-bacterial immunity. Int. J. Med. Microbiol. (2014), http://dx.doi.org/10.1016/j.ijmm.2014.06.010

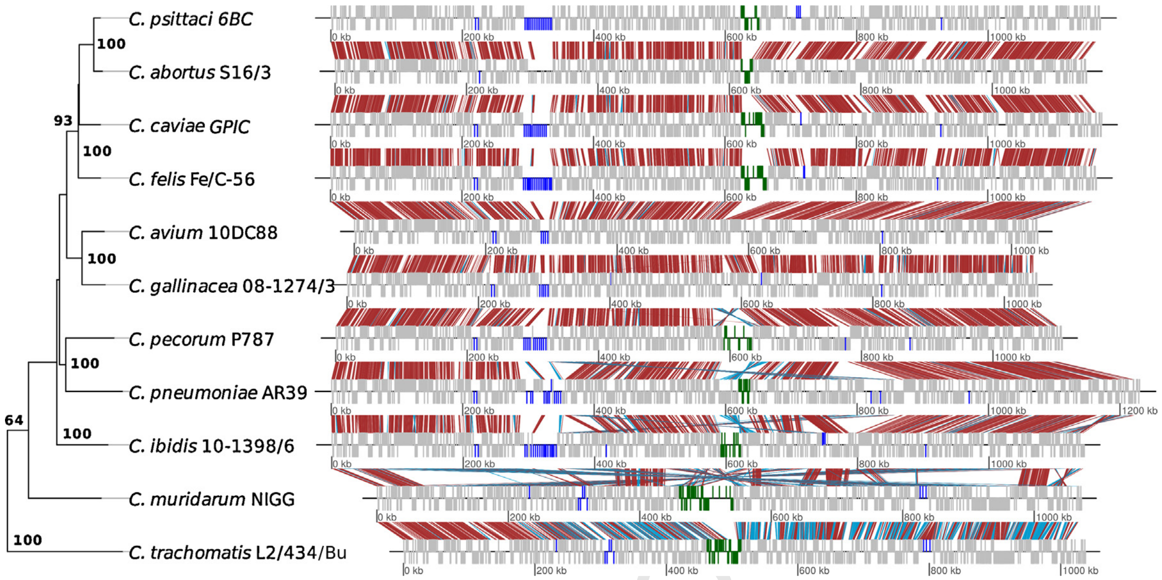


Fig. 1. Comparative genomics. Maximum-likelihood tree based on a random sample of 100 conserved orthologous genes and global comparison of synteny among the eleven fully sequenced species of the family *Chlamydiaceae* (*C. ibidis* is a taxon with *Candidatus* status). Gray tick marks above and below the sequence lines represent predicted CDSs on the plus and minus strands of the genomes, respectively. Color-marked are (blue) members of the polymorphic membrane protein (Pmp) family and (green) genes within the chlamydial plasticity zone (PZ). Red lines connecting genomes represent direct orthologous matches based on best reciprocal BLAT hits; blue lines represent reverse-complemented matches. Bootstrap values are based on 1000 resampled trees.

The size and organization of the PZ differs substantially among *Chlamydiaceae*, ranging from 45 genes in *C. trachomatis* to six genes in koala strain LPCoLN of *C. pneumonia* (Voigt et al., 2012). The PZ of *C. psittaci* is remarkable for its inclusion of a large intact toxin/adhesin gene similar to the EHEC adherence factor, as well as a *guaAB-add* cluster, which plays a role in salvaging biosynthesis of purine nucleotides necessary for bacterial growth. These genes are shared with *C. caviae* and *C. felis*, but are absent (or predicted to have loss-of-function frameshifts) in other chlamydial species, including the closely related *C. abortus* (Voigt et al., 2012). Interestingly, a recent comparative analysis of 20 *C. psittaci* strains revealed that only a subset of *C. psittaci* strains, the monophyletic"6BC clade", and its closest relatives may have acquired these genes through a single event of horizontal gene transfer, or they may have been lost from or degraded in more basal clades of *C. psittaci* on three independent occasions (Read et al., 2013). The 6BC clade comprises 10 strains including type strain 6BC/30 isolated from a parrot (Grinblat-Huse et al., 2011), and contains strains of both avian and mammalian origin corresponding to serotypes A and B (Read et al., 2013). In contrast, two further clades including serotypes C and D comprise only avian hosts. The 6BC clade shows extremely little internal variability (483 polymorphic nucleotide positions) coupled with a distinct lack of host specificity (Read et al., 2013). Interestingly, significant sequence differences between 6BC and non-6BC clades were observed in the large toxin locus. It is tempting to speculate that these remarkable differences among *C. psittaci* strains may have conferred adaptive advantages that facilitated host range expansion in the 6BC lineage.

In several chlamydial species, including *C. psittaci*, *pmp* genes have undergone a significant proliferation resulting in unusually high levels of mutational change within and across species. The Pmps are a large protein family, probably unique to the chlamydiae, with low overall amino acid similarity, but sharing multiple, often alternating motifs GGA (ILV) and FxxN. They are considered to be important in EB adhesion to host cells, molecular transport, and cell wall-associated functions (Longbottom et al., 1998), and have been identified as a critical factor in chlamydial pathogenesis (Tan et al., 2006). Pmps cluster phylogenetically into six subfamilies and are present in varying numbers ranging from a total of nine genes in *C. trachomatis* and *C. muridarum* to 21 in *C. pneumoniae* and *C. psittaci* (Voigt et al., 2012). Subfamily G/I is the largest and most rapidly evolving representative (Voigt et al., 2012). The tendency to proliferation of G/I family Pmps is especially pronounced in *C. psittaci* and its closest relatives, as well as in *C. pneumoniae*. While there are only two G/I *pmp* genes present in *C. trachomatis* and *C. muridarum*, there are 14 G/I family *pmp* genes present in the *C. psittaci* genome, several of which are unique to this species (Voigt et al., 2012). It has been suggested that the diversity in the *pmp* family of genes may have evolved in response to demands of the host environment, e.g. to escape systemic host response or adapt to multiple host environments (Tan et al., 2006). Thus, the extremely high variability of *pmp* genes in *C. psittaci* may also be related to its broad host spectrum.

While homologous recombination is a fundamental mechanism for genetic diversification in free-living bacteria, it is often thought to be limited in obligate intracellular bacteria, whose lifestyle should rarely allow encounters with sources of foreign DNA. With many genomes from the same species becoming available, it has become possible to apply traditional population genetics methods to whole genomes in order to test predictions as above with high precision. Thus, analysis of a large number of *C. trachomatis* genomes recently provided compelling evidence for relatively frequent homologous recombination in this species, despite the ecological barriers to contact posed by an intracellular lifestyle (Joseph et al., 2012). Similar evolutionary dynamics also play a surprisingly strong role in *C. psittaci*, where recombination events are spread evenly across the genome and were estimated to occur at roughly 1/20th the rate of mutation events. However, although recombination was less frequent than mutation, it accounted for three times more substitutions in a genome, because each recombination event affected several nucleotides at a time (Read et al., 2013). Strikingly, BLAST analysis of predicted recombinant segments indicated that roughly 50% of putative DNA imports derive from *C. abortus* or as-yet unknown/unsampled *C. psittaci* strains. This suggests that there is still substantial diversity in *C. psittaci* to be uncovered at the genome level.

UNCORRECTED PROOF

Please cite this article in press as: Knittler, M.R., et al., *Chlamydia psittaci*: New insights into genomic diversity, clinical pathology, host–pathogen interaction and anti-bacterial immunity. Int. J. Med. Microbiol. (2014), http://dx.doi.org/10.1016/j.ijmm.2014.06.010

Clinical pathology: the course of *C. psittaci* respiratory infection in a natural non-avian host

Animal models closely mimicking the relevant characteristics of human disease are still a key element for evaluating pathophysiological consequences and for understanding the underlying mechanisms of disease pathogenesis. Moreover, a valid animal model is a prerequisite for development and efficacy testing of therapeutics and vaccines.

With respect to research on respiratory *C. psittaci* infections, the bovine species offers a number of advantages stated in detail elsewhere (Reinhold et al., 2012). Morphological peculiarities of the bovine lung, including segmental anatomy and the lack of collateral airways, mirror pathophysiological mechanisms of pulmonary dysfunctions more clearly. The availability of a range of highly sensitive, non-invasive pulmonary function tests known from human medicine and previously validated for calves enables detailed evaluation of nature and scope of pulmonary disorders (Ostermann et al., 2014). Additionally, the body size of calves facilitates repeated sampling of biological specimens (e.g. blood and swabs), and thus follow-up monitoring of clinical and immunological parameters (see also Section "Local immune response: Development and resolution of pulmonary lesions after *C. psittaci* infection of calf and mouse"). A calf model using a bovine *C. psittaci* strain potentially allows detailed monitoring of transmission routes, molecular and physiological processes in the course of infection, as well as evaluation of a strain's virulence. The data obtained should be of interest for researchers in both veterinary and human medicine. In a recent series of infection trials, we were able to demonstrate and verify the essential characteristics of a *C. psittaci* infection in its various manifestations (Ostermann et al., 2013a,b, 2014; Reinhold et al., 2012).

First of all, the severity of disease during the acute phase of infection (Fig. 2A) proved clearly dose dependent. Respiratory and clinical signs, which peaked 2–3 days post-inoculation (dpi), were mild after inoculation of 10^6 inclusion-forming units (ifu) of *C. psittaci* strain DC15 per calf, moderate with doses of 10^7–10^8 ifu and severe for 10^9 ifu. This dose–response relation was highly reproducible in extent and quality of lung lesions, and also reflected in respiratory functions affecting pulmonary gas exchange (Ostermann et al., 2013b; Reinhold et al., 2012). Baseline effects due to chlamydial LPS or cell culture medium were ruled out by using controls that received UV-inactivated chlamydiae or medium, respectively.

Secondly, we were able to reproduce the whole variety of characteristic features that are observed in animals and humans affected by natural chlamydial infections. In the case of calves intrabronchially inoculated with 10^8 ifu/calf, a typical course included initial acute clinical illness (2–3 dpi), which subsided considerably, but not completely, within 10 dpi (Ostermann et al., 2013a). Subsequently, a protracted clinically silent course was indicated by intermittent mild symptoms, fecal pathogen excretion, transient chlamydemia, and slightly elevated levels of monocytes and lipopolysaccharide-binding protein (LBP) in blood. Interestingly, these phenomena were also observed as a consequence of naturally acquired, initially mild infection, as induced in a group of naïve calves (sentinels) that socialized with the acutely diseased animals. Again, this underlines the validity of the present bovine infection model and its potential to emulate the natural circumstances. The fact that neither group of calves completely eliminated the chlamydiae within five weeks after exposure (Ostermann et al., 2013a) is also in line with observations in animal herds, where (permanent or intermittent) carriers of chlamydiae represent a permanent reservoir for recurrent outbreaks of disease (Reinhold et al., 2011b). Under field conditions, even asymptomatic chlamydial infections in cattle herds are known to be associated with numerous negative effects on health and performance including significantly reduced growth rate in calves (Jaeger et al., 2007; Poudel et al., 2012; Reinhold et al., 2008).

Thirdly, the humoral immune response was generally weak (see also Section "Local immune response: Development and resolution of pulmonary lesions after *C. psittaci* infection of calf and mouse"). Only two-thirds of calves experimentally challenged with 10^8 ifu developed specific antibodies against *C. psittaci*as detected by immunoblotting. The onset of the antibody response occurred between 7 and 14 dpi. In sentinels, no humoral immune response was observed (Ostermann et al., 2013a; Reinhold et al., 2012). Although innate and acquired humoral immunity are also critically involved in the clearance of chlamydiae, this supports the notion that the cellular immune response plays a central role in controlling anti-bacterial immunity in infected hosts (see also Section "The cellular immune response: New insights into MHC I-mediated presentation of chlamydial antigens") and also imposes basic constraints on disease control approaches based solely on classical immunoprophylaxis.

Fourthly, the pathological manifestations of acute respiratory *C. psittaci* infection included inflammatory cells (mainly neutrophil granulocytes) recruited onto the site of infection (Figs. 2B and 3B; see also Sections "Local immune response: Development and resolution of pulmonary lesions after *C. psittaci* infection of calf and mouse" and "The innate immune response and its interaction with the adaptive response: Chlamydial infection and the complement system"). Pulmonary inflammation resulted in damage of the alveolar-capillary barrier, as indicated by altered cytology, elevated concentrations of eicosanoids and total protein in broncho-alveolar lavage fluid (Reinhold et al., 2012). Inflammation and the accumulation of detritus and protein-rich fluid caused (i) ventilatory disorders due to both airway obstructions and pulmonary restrictions and (ii) inhibited pulmonary gas exchange. Among the resulting functional consequences were reduced alveolar ventilation (hypoventilation) and reduced oxygen supply to the blood (hypoxemia). Attempts of compensation included the elevation of both respiratory rate and minute ventilation, clinically visible as dyspnea (Ostermann et al., 2013b, 2014).

Pulmonary dysfunctions were accompanied by metabolic disorders in the acid–base equilibrium revealing additional systemic consequences of an acute *C. psittaci* infection on blood pH and blood homeostasis in the host (Ostermann et al., 2014).

Taken together, the bovine model of respiratory *C. psittaci* infection revealed unprecedented insights into the very complex interactions between pathological changes, pathophysiological consequences and clinical signs induced by this pathogen in the mammalian host (Fig. 2B). With respect to comparative and translational medicine, this model builds a useful bridge between basic molecular research and clinical work in patients, both humans and animals. The elucidated host–pathogen interactions provide a suitable basis for further research focusing on new strategies to control this infection through treatment or vaccination.

Local immune response: development and resolution of pulmonary lesions after *C. psittaci* infection of calf and mouse

In recent studies, pulmonary lesions were investigated in models of experimental aerogenous infection with *C. psittaci* in calves (see Section "Clinical pathology: The course of *C. psittaci* respiratory infection in a natural non-avian host") and mice (see Sections "The innate immune response and its interaction with the adaptive response: Chlamydial infection and the complement system" and "The cellular immune response: New insights into MHC I-mediated presentation of chlamydial antigens") to investigate

UNCORRECTED PROOF

Please cite this article in press as: Knittler, M.R., et al., *Chlamydia psittaci*: New insights into genomic diversity, clinical pathology, host–pathogen interaction and anti-bacterial immunity. Int. J. Med. Microbiol. (2014), http://dx.doi.org/10.1016/j.ijmm.2014.06.010

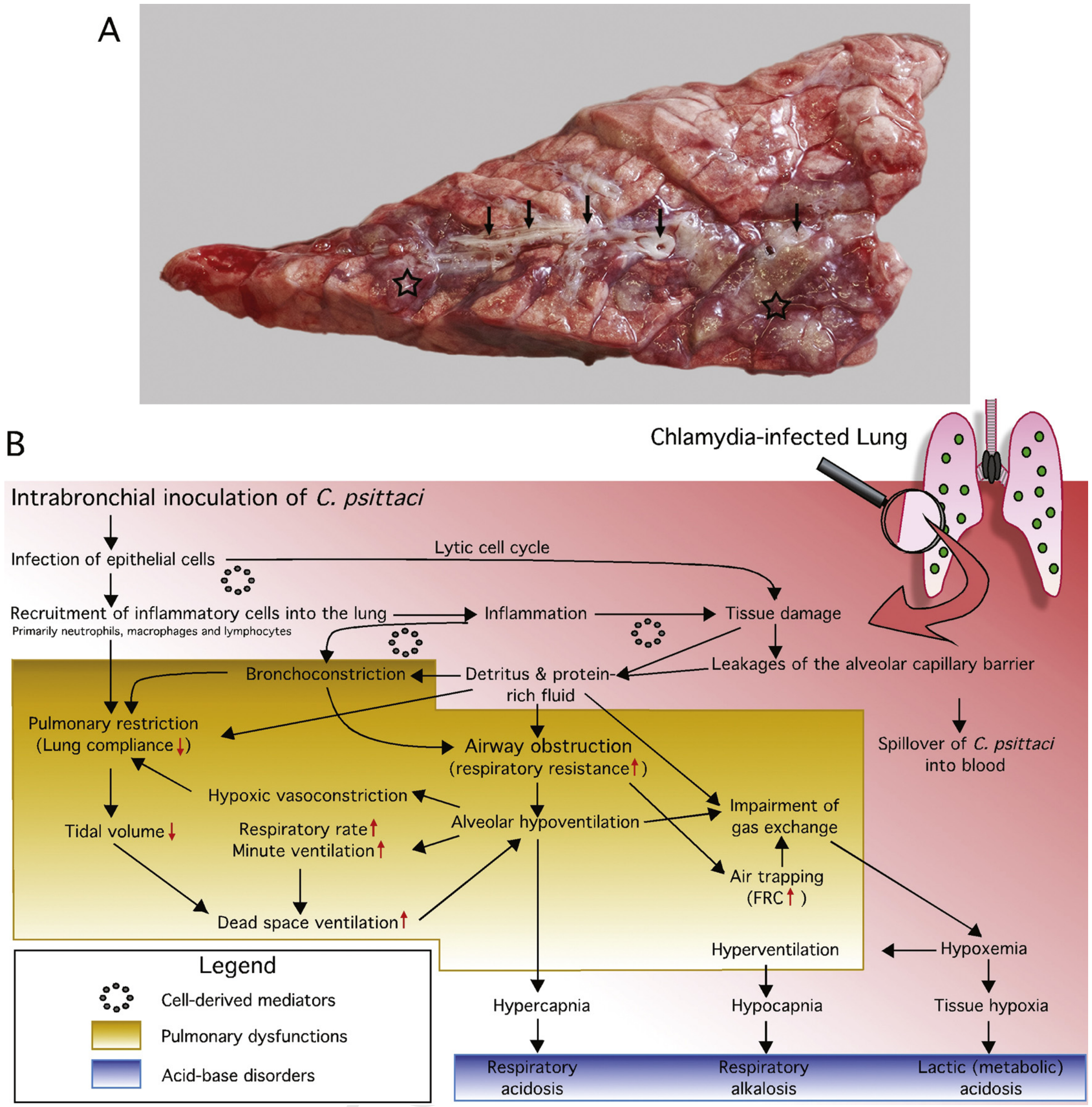

A
B
Chlamydia-infected Lung
Intrabronchial inoculation of *C. psittaci*
Infection of epithelial cells
Lytic cell cycle
Recruitment of inflammatory cells into the lung
Primarily neutrophils, macrophages and lymphocytes
Inflammation
Tissue damage
Leakages of the alveolar capillary barrier
Bronchoconstriction
Detritus & protein-rich fluid
Pulmonary restriction (Lung compliance ↓)
Airway obstruction (respiratory resistance ↑)
Spillover of *C. psittaci* into blood
Hypoxic vasoconstriction
Alveolar hypoventilation
Impairment of gas exchange
Tidal volume ↓
Respiratory rate ↑
Minute ventilation ↑
Air trapping (FRC ↑)
Dead space ventilation ↑
Hyperventilation
Hypoxemia
Legend
Cell-derived mediators
Pulmonary dysfunctions
Acid-base disorders
Hypercapnia
Hypocapnia
Tissue hypoxia
Respiratory acidosis
Respiratory alkalosis
Lactic (metabolic) acidosis

Fig. 2. Respiratory *Chlamydia psittaci* infection: complexity of pathogenetic mechanisms, pathophysiological interactions and systemic consequences in the mammalian host (bovine model). (A) Pulmonary lesions in the bovine lung. The section through a cranial lobe reveals several lobules with fibrinopurulent bronchopneumonia (asterisks) in close association with bronchi (arrows) in a calf 4 days after experimental inoculation with *C. psittaci*. (B) Inoculation of *C. psittaci* resulted in infection and replication of the pathogen in pulmonary epithelial cells and initiated an acute-phase reaction. Pro-inflammatory chemokines activated the inflammation cascade with recruitment and accumulation of phagocytic and reactive immune cells from the blood into the lung. Tissue damage and leakages of the alveolo-capillary membrane enabled a spillover of the pathogen into the bloodstream (chlamydemia). Within the lung, fluid, cells and debris in the alveoli and the interstitial tissue provided a barrier for oxygen transfer leading to oxygen deficiency in arterial blood (hypoxemia). Additional interacting factors induced by inflammatory mediators diminished multiple pulmonary functions. Analysis of respiratory mechanics revealed restriction of pulmonary tissue (i.e. limitations in lung compliance) and airway obstructions, both contributing to a reduction in alveolar ventilation (alveolar hypoventilation). Regarding the breathing pattern, expiration was more impaired than inspiration, which resulted in 'apped air', i.e. larger end-expiratory gas volume in the alveolar spaces due to incomplete deflation of the lung (i.e. increased functional residual capacity; FRC). The latter worsened the already existing alveolar hypoventilation and the already reduced oxygen delivery to the blood. In hypoventilated areas of the lung, the corresponding pulmonary arteries constricted (hypoxic pulmonary vasoconstriction; HPV), and those stiffer blood vessels with increased vascular resistance deteriorated the loss of compliant pulmonary properties. In the less compliant lung, the volume inhalable and exhalable per breath (tidal volume) was reduced. Due to the lack of oxygen in the blood, however, respiration was stimulated resulting in a compensatory increase of the respiratory rate and the volume of minute ventilation. Clinically, short, rapid breathing cycles appeared that were characterized by an increased percentage of dead space ventilation and a reduced percentage of alveolar ventilation. Tissue damage and pulmonary dysfunctions had a strong impact on acid–base equilibrium. While alveolar hypoventilation may result in accumulation of CO_2 in the body (hypercapnia) and a consecutive respiratory acidosis (blood pH ↓), the attempt to compensate hypoxemia by a dramatically intensified respired volume per time (hyperventilation) was associated with a loss of CO_2 (hypocapnia) driving blood pH in the direction of respiratory alkalosis (pH ↑). In addition, the demand of oxygen in peripheral tissues in relation to a reduced oxygen supply by the blood stimulated anaerobe metabolism (increased production of L-lactate and protons), which influenced blood pH in the direction of metabolic acidosis (pH ↓). Additional complex metabolic influences on blood pH are described in Ostermann et al. (2014). In general, numerous acidic effects were counterbalanced by a variety of alkalotic effects influencing the acid–base equilibrium in the host organism in multiple ways. In comparative and translational medicine, this integrated large animal model has fundamentally improved our understanding of pathogenetic mechanisms and complex pathophysiological interactions during respiratory *C. psittaci* infection.

species-specific reaction patterns to *C. psittaci* (Fiegl et al., 2013; Lambertz, 2011; Möhle, 2011; Reinhold et al., 2012).

In calves, intra-tracheal application of viable *C. psittaci* induced pneumonia in all animals inoculated (Fig. 2), but in none of the controls (Reinhold et al., 2012). In addition to the dose-dependent clinical score (see Section "Clinical pathology: The course of *C. psittaci* respiratory infection in a natural non-avian host"), the type of lesions changed with increasing dose of inoculum from purulent to fibrinous to necrotizing. Thus, the type of acute lesions differed distinctly from the interstitial and atypical pneumonia generally

Please cite this article in press as: Knittler, M.R., et al., *Chlamydia psittaci*: New insights into genomic diversity, clinical pathology, host–pathogen interaction and anti-bacterial immunity. Int. J. Med. Microbiol. (2014), http://dx.doi.org/10.1016/j.ijmm.2014.06.010

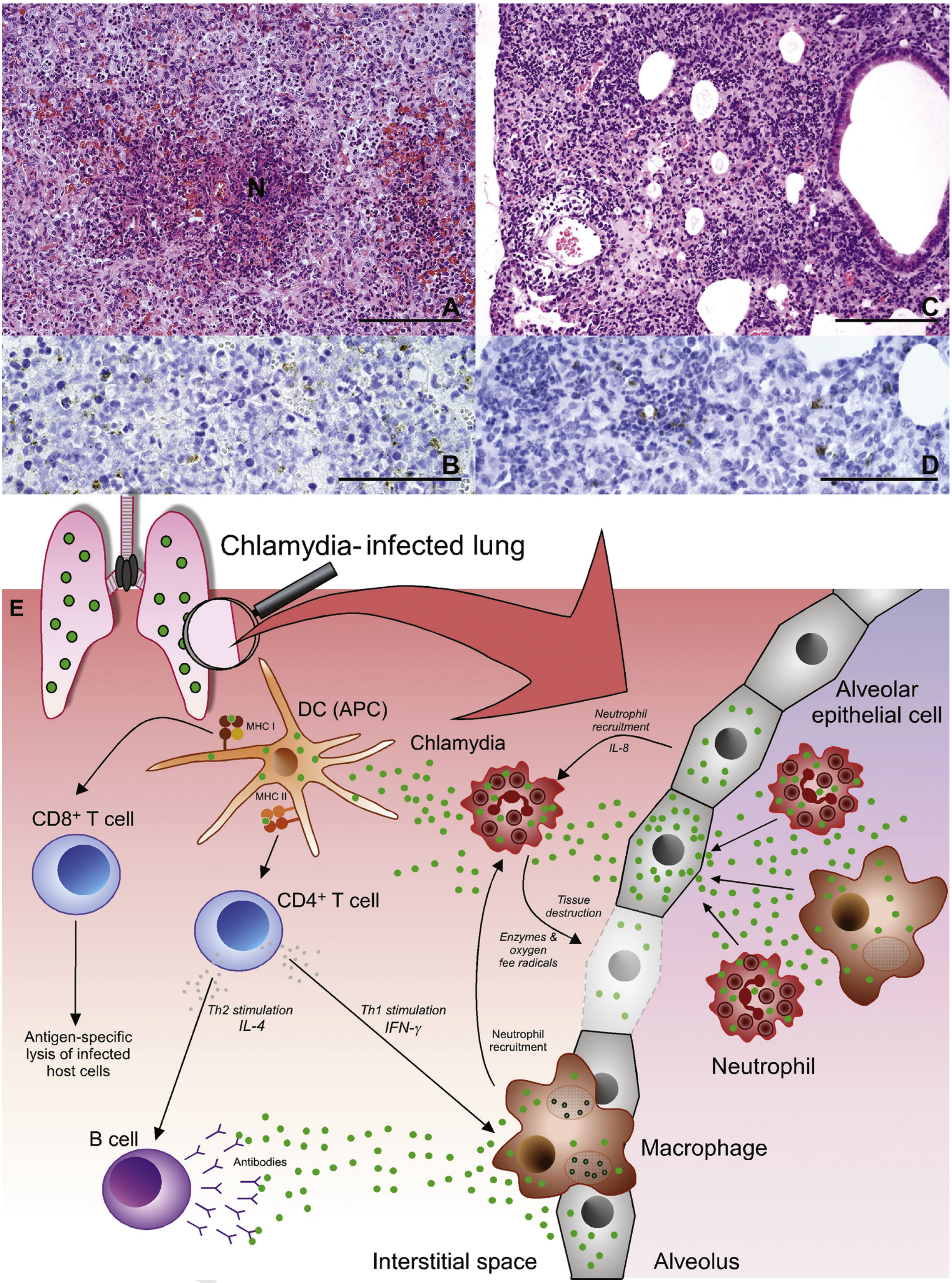

N
A
C
B
D
Chlamydia- infected lung
E
DC (APC)
MHC I
MHC II
Chlamydia
Neutrophil recruitment
IL-8
Alveolar epithelial cell
CD8⁺ T cell
CD4⁺ T cell
Tissue destruction
Enzymes & oxygen fee radicals
Th2 stimulation
IL-4
Th1 stimulation
IFN-γ
Neutrophil recruitment
Antigen-specific lysis of infected host cells
Neutrophil
B cell
Antibodies
Macrophage
Interstitial space
Alveolus

Fig. 3. Pulmonary lesions (A–D) and model of local immune response to Chlamydiae (E). Pulmonary lesions 4 days after experimental inoculation of a calf (A and B) and a mouse (C and D) with *C. psittaci*. Lesions in the calf (A) are characterized by fibrinopurulent exudate and foci of necrosis (N), whereas neutrophilic infiltrates predominate in the mouse (C). Numerous chlamydial inclusions (brown) are present in intra-lesional neutrophils and macrophages of calf (B) and mouse (D). A, C bars = 100 μm. B, D bars = 200 μm. In the model of local immune response to chlamydiae (E), infected alveolar epithelial cells secrete interleukin-8 (IL-8), which is a potent pro-inflammatory cytokine playing a key role in recruitment and activation of neutrophils. Neutrophils produce enzymes and reactive oxygen species to kill bacteria and provide the first line of defense against chlamydiae together with macrophages and dendritic cells (DCs). Binding of conserved pathogen-derived molecules by specific receptors on the cell

Please cite this article in press as: Knittler, M.R., et al., *Chlamydia psittaci*: New insights into genomic diversity, clinical pathology, host–pathogen interaction and anti-bacterial immunity. Int. J. Med. Microbiol. (2014), http://dx.doi.org/10.1016/j.ijmm.2014.06.010

UNCORRECTED PROOF

reported in calves and humans (Beeckman and Vanrompay, 2009; Jee et al., 2004).

The time course of infection was followed in inoculated calves from 2 days to 1 month after inoculation (Lambertz, 2011; Möhle, 2011). The pneumonic exudate predominantly consisted of neutrophil granulocytes (see also Sections "Clinical pathology: The course of *C. psittaci* respiratory infection in a natural non-avian host" and "The innate immune response and its interaction with the adaptive response: Chlamydial infection and the complement system") at 2 and 3 dpi. Lesions were most severe at days 3 and 4 pi with fibrinous exudate, areas of necrosis and increasing numbers of alveolar macrophages (Fig. 3A). Chlamydial inclusions were detected initially in alveolar epithelial cells and then increasingly in neutrophil granulocytes and alveolar macrophages especially in areas of necrosis (Fig. 3B). Ultrastructural investigation revealed infection of alveolar epithelial cells, neutrophil granulocytes and alveolar macrophages with *C. psittaci*, but multiplication was seen in alveolar epithelial cells only. At day 7 pi, first signs of regeneration were seen: necrotic pulmonary tissue was demarcated by proliferating alveolar epithelial cells and dense infiltrates of macrophages. The zones of regeneration became confluent at day 10 pi and were infiltrated by dendritic cells (DCs) (see also Section "The cellular immune response: New insights into MHC I-mediated presentation of chlamydial antigens"), macrophages, T and B lymphocytes and plasma cells. At day 14 pi, the numerous lymphocytes, macrophages, DCs and ectopic lymphoid follicles in the lesions indicated a strong local immune response that caused a marked reduction of the number of chlamydiae. The pulmonary tissue had regenerated one month pi except for a few small foci of fibrous scar tissue. Number and distribution of unorganized immune cells were comparable to control calves, but bronchus-associated lymphoid tissue was more pronounced. The intrapulmonary infection of calves with *C. psittaci* allows a detailed study of the steps hypothesized to be involved in the development and organization of lung damage in animals and humans (Beers and Morrisey, 2011).

For the zoonotic agent *C. psittaci*, knowledge about shedding and potential routes of transmission is important. The detection of *C. psittaci* in nasal mucosa and trachea from 2 to 4 days pi was interpreted as excretion of the pathogen with infected exudate from the lung, since chlamydial inclusions were detected in luminal exudate in the trachea by immunohistochemistry. The excretion was limited to the short time period of severe lesions and to low numbers of chlamydial inclusions in the exudate (Lambertz, 2011). Dissemination of *C. psittaci* from the infected lung to sites that are considered to be preferential targets for chronic chlamydial diseases in humans, e.g. aorta and synovial membranes, was detected by PCR. The lack of tissue reaction in these target tissues argues against an active infection and confirms findings from infection experiments in rabbits with *C. pneumonia* (Gieffers et al., 2004).

In another experimental setting, mice were intranasally inoculated with the same strain of *C. psittaci* that was used in calves (see Sections "The innate immune response and its interaction with the adaptive response: Chlamydial infection and the complement system" and "The cellular immune response: New insights into MHC I-mediated presentation of chlamydial antigens"). The initial phase of infection was similar in both species. At day 4 pi, differences in reaction of mice versus calves became apparent, i.e. extensive purulent inflammation with increasing numbers of macrophages was found in mice (Fig. 3C), while calves developed severe fibrinous exudation with tissue destruction (Fig. 3A). Large numbers of chlamydiae were associated with neutrophil granulocytes and macrophages in the lesions of both hosts (Fig. 3B and D). The different types of inflammatory lesions resulted in distinct healing processes, i.e. mice developed lymphohistiocytic bronchointerstitial pneumonia, whereas regeneration with alveolar epithelial type 2 cell hyperplasia occurred in calves. Both mounted a massive local immune response that eliminated the chlamydiae (see also Sections "The innate immune response and its interaction with the adaptive response: Chlamydial infection and the complement system" and "The cellular immune response: New insights into MHC I-mediated presentation of chlamydial antigens").

Based on the morphological findings, the following hypothesis for the pathogenesis of pulmonary *C. psittaci* infection can be suggested (Fig. 3E). There is an initial infection of alveolar epithelial cells with *C. psittaci*. The multiplication of *C. psittaci* is followed by a rapid influx of neutrophil granulocytes most likely mediated by cytokines released from infected cells (see Sections "The innate immune response and its interaction with the adaptive response: Chlamydial infection and the complement system" and "The cellular immune response: New insights into MHC I-mediated presentation of chlamydial antigens"). These initial steps of the inflammatory reaction are consistent with the "cellular paradigm" in chlamydial infections (Stephens, 2003). Degranulation and decay of neutrophil granulocytes cause extensive species-specific damage of the pulmonary tissue. Progressive elimination of *C. psittaci* by adaptive immune response including DCs (see Section "The cellular immune response: New insights into MHC I-mediated presentation of chlamydial antigens"), T and B cells and regeneration of the pulmonary tissue follow. These findings go beyond the results from the large number of studies where only lesions at a single time point after their initiation or during their development were described. By focusing not only on the pathogen, but also the host's local immune response, our study allows insights into host–pathogen interactions at the commencement, progression and regression of the infection.

Host–pathogen interaction: identifying the molecular partners on both sides of chlamydial cell infection

During the biphasic developmental cycle replicating bacteria acquire energy and biosynthetic precursors from the infected cell. Chlamydiae are capable of modulating cellular functions, such as apoptotic programs and immune response (see also Sections "The innate immune response and its interaction with the adaptive response: Chlamydial infection and the complement system" and "The cellular immune response: New insights into MHC I-mediated presentation of chlamydial antigens") (reviewed by Bastidas et al., 2013). Studies on inhibitors of bacterial protein synthesis suggest that modulation of host cell functions requires the activity of specific chlamydial proteins (see also Section "Host adaptation of *C. psittaci*: Escape from the avian innate immune response").

C. trachomatis dramatically rearranges the architecture of the eukaryotic host cell including the Golgi apparatus (GA), localization and appearance of the microtubule-organizing center (MTOC) and polarization of migrating cells (Heuer et al., 2009; Heymann et al., 2013; Johnson et al., 2009; Knowlton et al., 2011). These changes contribute either directly to bacterial growth or have implications on pathogenesis. Currently, it is unknown whether these observed changes are specific for *C. trachomatis* or if zoonotic *Chlamydia* strains including *C. psittaci* can induce similar changes. We have experimentally addressed this question by comparing the GA structure in a human cell culture model infected either with *C. psittaci*

surface of macrophages and DCs activates the production of pro-inflammatory cytokines and chemokines, which help to attract other effector cells to the site of infection. Chlamydia-activated DCs present bacteria-derived antigens to T cells and promote the differentiation of naive T cells to various subtypes of effector CD4+ and CD8+ T cells. CD4+ T helper 1 (Th1) cells secrete IFN-γ, which activates the anti-microbial activity of macrophages and helps B-cell production of chlamydia-specific antibodies, whereas Th2 cells provide help for B-cell production of antibodies.

Please cite this article in press as: Knittler, M.R., et al., *Chlamydia psittaci*: New insights into genomic diversity, clinical pathology, host–pathogen interaction and anti-bacterial immunity. Int. J. Med. Microbiol. (2014), http://dx.doi.org/10.1016/j.ijmm.2014.06.010

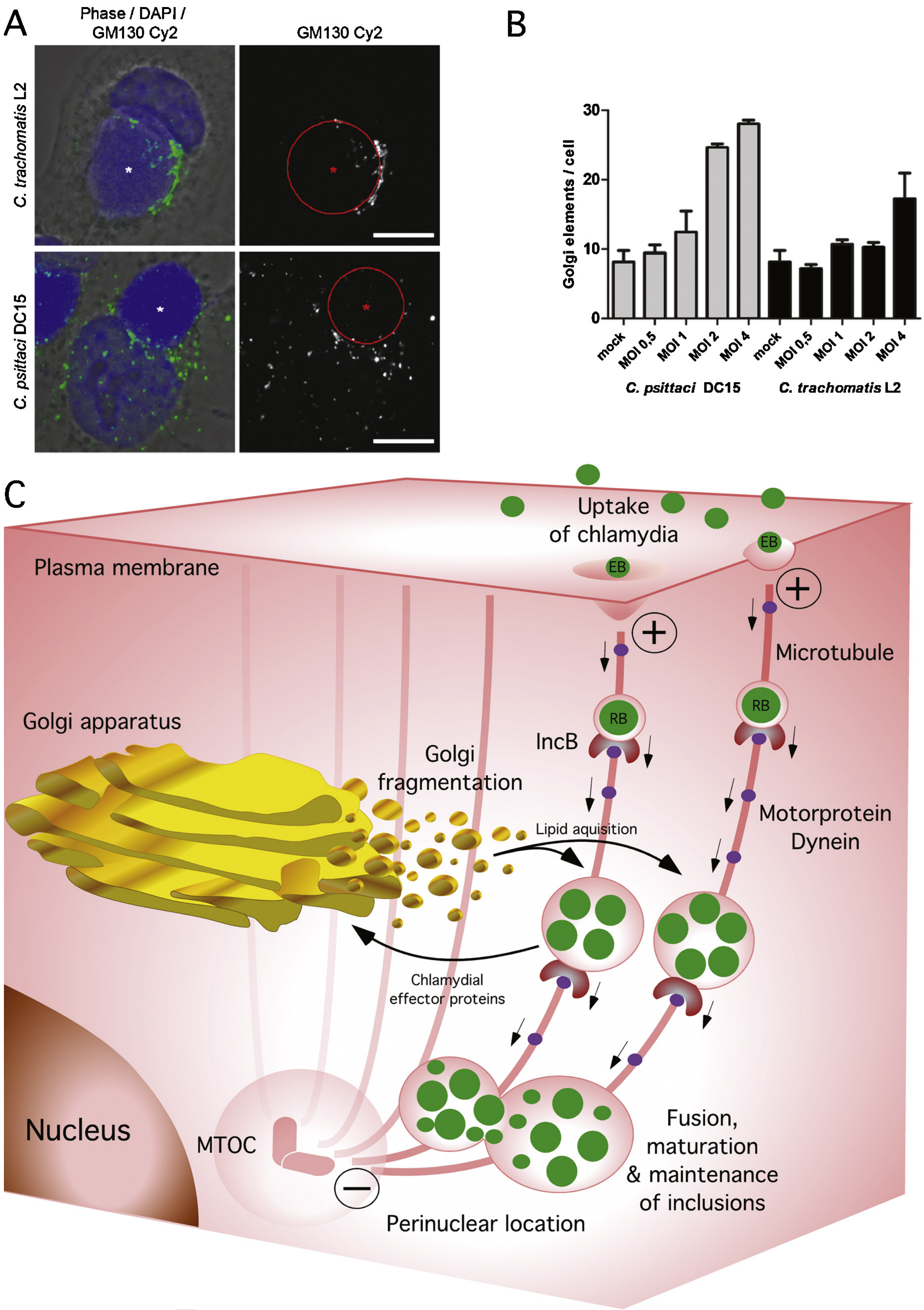

A
Phase / DAPI / GM130 Cy2
GM130 Cy2
C. trachomatis L2
C. psittaci DC15
B
Golgi elements / cell
30
20
10
0
mock
MOI 0,5
MOI 1
MOI 2
MOI 4
mock
MOI 0,5
MOI 1
MOI 2
MOI 4
C. psittaci DC15
C. trachomatis L2
C
Uptake of chlamydia
EB
EB
Plasma membrane
Microtubule
RB
RB
IncB
Golgi apparatus
Golgi fragmentation
Motorprotein Dynein
Lipid aquisition
Chlamydial effector proteins
Nucleus
MTOC
Fusion, maturation & maintenance of inclusions
Perinuclear location

Fig. 4. Human and zoonotic chlamydial infections result in Golgi fragmentation in human cell lines. (A) HeLa cells were infected with either *C. trachomatis* L2 or *C. psittaci* at MOI 2. At 26 h p.i., cells were fixed and stained for the Golgi matrix protein, GM130 (in green). Cells were counter-stained using DAPI and analyzed at a Zeiss LSM 780. Overlay with a phase contrast is shown. Inclusions are indicated using a red line. Scale bar; 10 μm. (B) HeLa cells were infected with *C. trachomatis* L2 or *C. psittaci* using different MOIs. At 26 hpi, cells were fixed and the GA was visualized by GM130 staining. Golgi elements were then quantified using ImageJ. (C) Intracellular transport and lipid supply of chlamydial inclusion. Chlamydia exploits the host's intracellular trafficking machinery by recruiting the microtubule motor protein dynein to the outer surface of the vacuole (mediated by IncB), which drives the migration of inclusions toward the minus end of microtubules and the microtubule-organizing center (MTOC), where it

Please cite this article in press as: Knittler, M.R., et al., *Chlamydia psittaci*: New insights into genomic diversity, clinical pathology, host–pathogen interaction and anti-bacterial immunity. Int. J. Med. Microbiol. (2014), http://dx.doi.org/10.1016/j.ijmm.2014.06.010

or *C. trachomatis* (Fig. 4). Infections resulted in fragmentation of the GA into smaller Golgi elements, and GM-130-positive Golgi elements were recruited to *C. trachomatis* inclusions. In contrast, *C. psittaci*-infected cells showed GM-130-positive Golgi elements scattered throughout the cytosol with no obvious recruitment to the inclusions (Fig. 4A). Quantification of the number of Golgi elements demonstrated Golgi fragmentation is MOI dependent but more pronounced in the case of *C. psittaci* infections (Fig. 4B). In summary, Golgi fragmentation is a hallmark of human and zoonotic chlamydial infections at least in a human cell culture model but mode of fragmentation and/or trafficking of fragmented Golgi elements differ between species. These differences are likely reflected in divergence of chlamydial effector proteins.

All *Chlamydiaceae* spp. possess genes encoding core components of a Type III Secretion (TTS) apparatus, a protein transport system used by Gram-negative bacteria to translocate proteins into the cytoplasm of the host cell. Therefore, it is commonly accepted that chlamydial effector proteins are targeted by the TTS to the inclusion membrane and that their interaction with host proteins cause the modulation of host cell functions (Peters et al., 2007). Interactions of chlamydial effector proteins with host proteins seem to play a role at all stages of chlamydial development, i.e. from adhesion and internalization of EBs till their exit from the host cell.

An important group of chlamydial effector proteins are the Inc (inclusion) proteins, which share a common secondary structural feature of a bilobed hydrophobic domain. While this hydrophobic domain enables the anchoring of the proteins in the inclusion membrane the cytoplasmic tail is responsible for interaction with host proteins (Dehoux et al., 2011). By direct interaction with host cell proteins, Inc proteins may affect different cell functions including signaling and trafficking. One of the first discovered protein–protein interactions was the association of IncG with mammalian 14-3-3β protein (Scidmore and Hackstadt, 2001), which contributes to circumvention of host cell apoptosis. The Inc protein CP0236 impairs IL-17 signaling via interaction with human Act1 (Wolf et al., 2009). Furthermore, IncD of *C. trachomatis* was shown to interact with the lipid transfer protein CERT (Derre et al., 2011), while another Inc protein, CT229, associated with the GTPase Rab4 (Rzomp et al., 2006). Both interactions may affect trafficking processes of the host cell. Finally, Inc protein CT228 was found to recruit elements of the myosin phosphatase pathway to regulate release mechanisms (Lutter et al., 2013).

In a recent study, we focused on the IncA and IncB proteins of the zoonotic agent *C. psittaci* using yeast two-hybrid screens to search for interacting host proteins that could indicate functions of these proteins. In a first attempt, the interaction between type III-secreted protein IncA of *C. psittaci* and host protein G3BP1 was identified. In GST-pull down and co-immunoprecipitation experiments both in vitro and in vivo interaction between full-length IncA and G3BP1 were shown. G3BP1 harbors a phosphorylation-dependent RNase activity that specifically cleaves the 3′-untranslated region of human *c-myc* mRNA. Using fluorescence microscopy, the localization of G3BP1 near the inclusion membrane of *C. psittaci*-infected Hep-2 cells was demonstrated. Notably, infection of Hep-2 cells with *C. psittaci* and overexpression of IncA in HEK293 cells led to a decrease in c-Myc protein concentration. This effect could be ascribed to the interaction between IncA and G3BP1 since overexpression of an IncA mutant construct disabled to interact with G3BP1 failed to reduce c-Myc concentration. We hypothesize that lowering the host cell c-Myc protein concentration may be part of a strategy employed by *C. psittaci* to avoid apoptosis and scale down host cell proliferation (Borth et al., 2011). Furthermore, it was shown that IncB in early inclusions of *C. psittaci* interacts and co-localizes with components of the dynein complex (data not shown). It is known that within the first few hours of infection, endocytosed *C. trachomatis* EBs are transferred to a perinuclear location corresponding to the MTOC of the host cell using the tubulin–dynein system (Grieshaber et al., 2003). Like *C. trachomatis*, *C. psittaci* also utilizes dynein motor proteins for optimal early development (Escalante-Ochoa et al., 2000). The interplay of IncB of *C. psittaci* with dynein components is the first indication of chlamydial proteins being involved in this transport process and possibly affecting the localization and structural appearance of the MTOC (Heuer et al., 2009) to support bacterial growth in the infected cell.

Host adaptation of *C. psittaci*: Escape from the avian innate immune response

Factors and mechanisms determining host specificity of *Chlamydia* spp. are still largely unknown. In a recent study (Braukmann et al., 2012), *C. psittaci* and *C. abortus* were compared for their invasiveness, virulence, and capability of eliciting an immune response in a chicken embryo model. This system has been used for decades, particularly to culture viruses and intracellular bacteria. As it involves inoculation onto the chorioallantoic membrane (CAM) the model resembles natural infection across epithelial layers. In fact it represents a valuable tool to investigate the innate immune response to chlamydial infection and to identify molecular processes on the pathogen's side in conjunction with the host's response that could explain differences in host adaptation between chlamydial species.

When compared with *C. abortus*, which is closely related but prefers non-avian hosts, *C. psittaci* displayed a significantly better capability of disseminating in host organs and elicited higher macrophage numbers and stronger expression of a sub-set of immune-related proteins. Nevertheless, the immune response patterns to both pathogens were largely similar, with expression of the pro-inflammatory mediators IL-1β, IL-6, IL-8, LITAF, IL-17 and iNOS, as well as IFN-γ, IL-12, IL-18, and TLR4, being significantly upregulated in CAM, liver and spleen of *C. psittaci*- and *C. abortus*-infected embryos. However, significant differences in favor of *C. psittaci* were seen in embryo mortality (Fig. 5) and mRNA expression rates of chlamydial genes, such as *incA*, *groEL* (in CAM, liver, spleen), *cpaf* and *ftsW* (in CAM). It is conceivable that elevated expression levels of the IncA protein would contribute to stabilization of the inclusion where the bacteria are residing (see also Section "Host–pathogen interaction: Identifying the molecular partners on both sides of chlamydial cell infection"). As expression of FtsW is used as an indicator of the pathogen's capability to replicate in the host, elevated mRNA expression rates of this gene could explain the better proliferation of *C. psittaci* throughout the infection. Furthermore, upregulation of the gene encoding the chlamydial heat shock protein GroEL (Hsp60) coincided with the elevated macrophage influx during *C. psittaci* infection. GroEL serves as a chaperone during intracellular development (Zugel and Kaufmann, 1999) and seems to be involved in the induction of immune response during human chlamydial infection, where it co-localizes with infiltrating macrophages (Kol et al., 1999).

Compared to *C. abortus*, *C. psittaci* appeared to be more efficient in actively suppressing the host immune response, which enabled better dissemination in the host. In this context, up-regulation of

UNCORRECTED PROOF

resides in a central perinuclear position to establish its intracellular niche near the Golgi apparatus. Fusion, maturation and maintenance of chlamydial inclusions require clustering around the perinuclear MTOC. During the maturation of inclusions chlamydial effector proteins cause Golgi fragmentation and thereby ensure lipid acquisition and bacterial growth.

Please cite this article in press as: Knittler, M.R., et al., *Chlamydia psittaci*: New insights into genomic diversity, clinical pathology, host–pathogen interaction and anti-bacterial immunity. Int. J. Med. Microbiol. (2014), http://dx.doi.org/10.1016/j.ijmm.2014.06.010

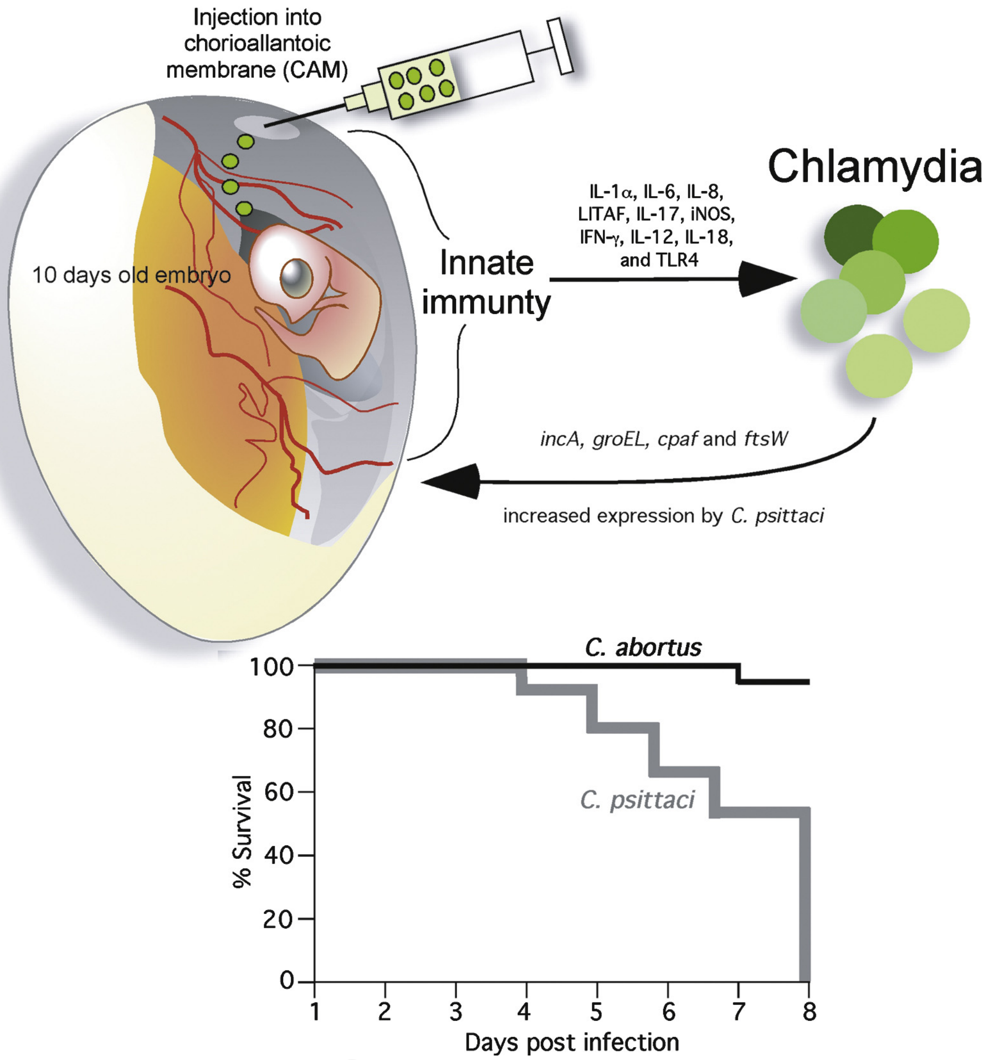


Fig. 5. Chlamydial infection in chicken embryos. Fertilized chicken eggs were incubated at 37.8 °C and a relative humidity of 60%. Inoculation of *C. psittaci* and *C. abortus* was conducted at embryonic developmental day (ED) 10. For infection, the bacterial suspension was transferred onto the chorioallantoic membrane (CAM (upper left panel)). The immune response patterns to both pathogens were largely similar, with expression of the pro-inflammatory mediators (upper right panel). Significant differences in favor of *C. psittaci* were seen mRNA expression rates of essential chlamydial genes (upper right panel), such as *incA*, *groEL*, *cpaf* and *ftsW* and in embryo mortality. Kaplan–Meier curves showing the survival of chicken embryos after infection with *C. psittaci* and, for comparison, *C. abortus* (lower panel). On ED 10, embryos ($n = 20$ per group) were inoculated with comparable amounts of chlamydiae (9×10^4–1×10^5 IFU/egg) and their vitality was determined by daily candling.

cpaf (see also Section “The cellular immune response: New insights into MHC I-mediated presentation of chlamydial antigens”) probably enabled the pathogen to influence the host reaction in its favor. The results indicate a relevance of *incA* (see also Section “Host-pathogen interaction: Identifying the molecular partners on both sides of chlamydial cell infection”), *groEL*, *cpaf* and *ftsW* for the pathogenicity of *C. psittaci*. The data of the in vivo study provide a basis for a more detailed insight and better understanding of host–pathogen interactions in the course of infection. The fact that *C. psittaci* coped far better than *C. abortus* with the onset of the avian embryo’s innate immune response by upregulating selected genes may be a key to understanding the mechanisms underlying host adaptation and etiopathology.

The innate immune response and its interaction with the adaptive response: Chlamydial infection and the complement system

The complement system consists of approximately 40 serum factors, mainly zymogens, their cleavage products and receptors. It is activated by LPS or lectins and other components of bacterial and viral surfaces. The activated complement system modulates inflammatory responses and protects against extracellular pathogens. Complement is often regarded as a key player of innate immunity (which can additionally be activated by immune complexes). However, there is increasing evidence that it can additionally trigger, enhance and orchestrate adaptive immune responses, at least in viral infections (reviewed by Klos et al., 2013). A complement system with homologous factors and identical function was found not only in mammals, but also in other vertebrate species including birds (Sharma, 1997).

Biologically active cleavage products of the circulating inert complement factors are locally released upon activation. Some of them become part of enzyme complexes, which permit successive steps in the cascade. Others remain soluble, bind to receptors on the host cell surface and act as extracellular mediators for intracellular signaling events. The three classical and broadly known effector functions of the complement cascade, i.e. opsonization via antigen-bound C3b, lysis by the membrane attack complex, pro-inflammation via the anaphylatoxins, are located downstream of complement factor C3. The anaphylatoxic peptides C3a and C5a, cleavage products of factor C3 and factor C5, respectively, activate

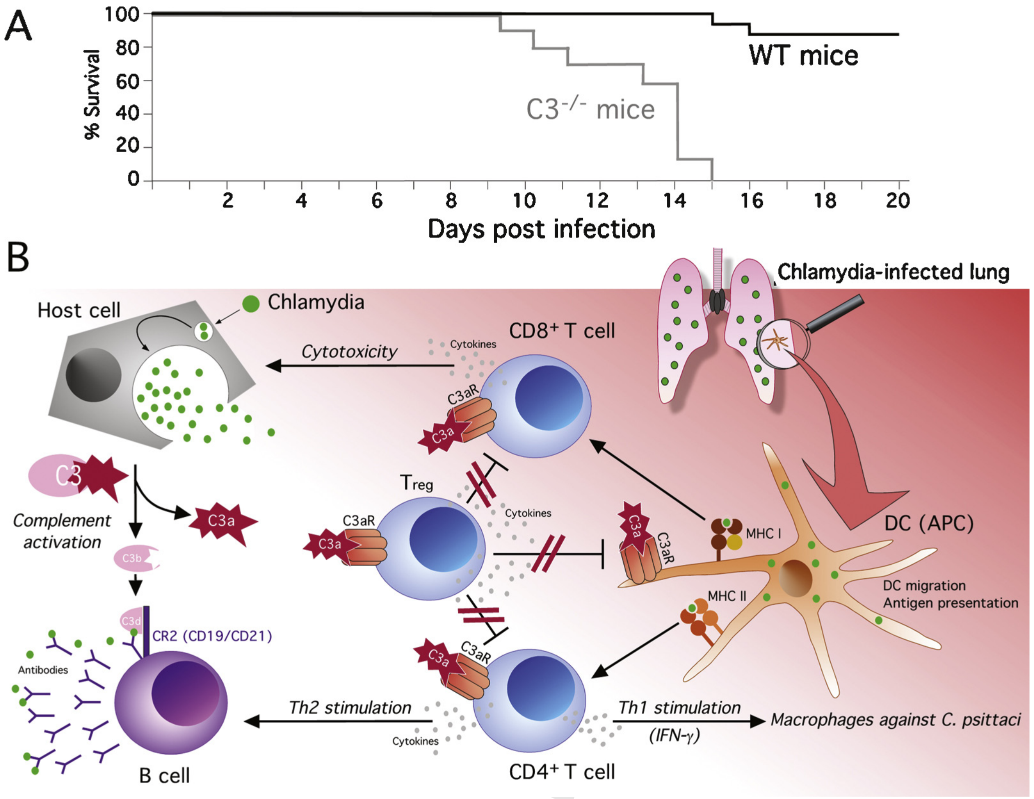


Fig. 6. (A) Complement factor C3 provides protection of mice against *C. psittaci*. $C3^{-/-}$ and the corresponding wild-type (WT) mice were intranasally infected with 10^4 ifu of *C. psittaci* strain DC15. Survival rates were determined for up to 21 days post-infection. (B) Assumed C3a/C3a-receptor (C3aR)-dependent regulation of the adaptive immune response against chlamydiae. Chlamydial elementary bodies (green) activate the complement cascade leading to the cleavage of inactive complement factor C3 into active C3a and C3b. Bound to its antigen, the C3b derivative C3d stimulates via complement receptor 2 (CR2) B cells and antibody production of plasma cells. The C3aR is expressed on mature dendritic cells (DCs) and, most likely, also on various subtypes of activated T cells including $CD4^+$ $CD25^+$ regulatory T cells (Treg cells). For the development of an effective adaptive protection against the bacteria, the immune-modulator C3a stimulates directly $CD4^+$ (Th1 and Th2) and cytotoxic $CD8^+$ T cells. Alternatively, C3aR-signaling down-regulates Treg cells thereby decreasing their inhibition of $CD4^+$ and $CD8^+$ T cells and DCs. Moreover, C3a/C3aR can activate DCs leading to their improved migration to the draining lymph nodes and thus, to an augmented presentation of chlamydial antigens to lymphocytes.

UNCORRECTED PROOF

their G-protein coupled receptors. C3a receptor (C3aR) and C5a receptor (C5aR, $C5a_1$ receptor, CD88) are expressed on cells of myeloic origin including granulocytes, monocytes/macrophages, DCs, mast cells, and even progenitor cells (see also Sections "Local immune response: Development and resolution of pulmonary lesions after *C. psittaci* infection of calf and mouse" and "The cellular immune response: New insights into MHC I-mediated presentation of chlamydial antigens"). Moreover, the anaphylatoxin receptors are also found on several non-myeloic cells and activated T cells. Albeit C3aR and C5aR are closely related their functions overlap only partially.

EBs of *C. trachomatis* activate complement in vitro thereby reducing their infectivity in cell culture (Fedorko et al., 1987; Lin et al., 1992; Megran et al., 1985). Except for this fact, there was only limited knowledge about the role of complement during infections with *Chlamydia* or other intracellular bacteria at the beginning of the present research project. In a mouse model of intranasal *C. psittaci* lung infection, an early, high and long-lasting activation of the complement system was found (Bode et al., 2012). Intriguingly, only after the first week of infection complement became highly protective with an almost 100-fold increased susceptibility of $C3^{-/-}$ compared to wild type (WT) mice (Fig. 6A). In contrast, there were only minor differences after *C. psittaci* infection of complement factor C5-deficient and $C5aR^{-/-}$ in comparison to WT mice. This indicates that the main protective function of the cascade locates downstream of C3, but upstream of C5. It also points strongly toward C3a/C3aR as promising candidates of combatants for an effective defense against this intracellular microorganism.

Indeed, in our most recent study using $C3aR^{-/-}$ mice, the hypothesis that the C3aR is indispensable for effective protection and improved survival in *C. psittaci* lung infection has been confirmed (Dutow et al., 2013). During the first week, the course of disease was almost identical in infected WT and $C3aR^{-/-}$ mice. However, from then on, the ability to respond via C3aR to C3a significantly shortened severe acute pneumonia, led to drastically higher survival rates, more weight gain and lower clinical score. The bacterial clearance from lung and spleen was impaired in the absence of C3aR. Inflammatory parameters in lung homogenate including several cytokines and neutrophil granulocytes (and myeloperoxidase as granulocyte marker) (see also Section "Local immune response: Development and resolution of pulmonary lesions after *C. psittaci* infection of calf and mouse") remained elevated in $C3aR^{-/-}$ mice, as determined 14 days p.i. The most likely explanation for the sustained inflammatory response is the ongoing *C. psittaci* infection and the higher bacterial burden in the absence of C3aR.

Furthermore, the most likely reason for the impaired clearance of chlamydiae and the high lethality of the $C3aR^{-/-}$ mice during the second and third weeks p.i. has been addressed The time course suggested C3a/C3aR-dependent effects on the adaptive immune response. Indeed, the size of the lung-draining lymph nodes of $C3aR^{-/-}$ mice increased less during chlamydial infection until day 9, i.e. the last time point before $C3^{-/-}$ and $C3aR^{-/-}$ animals can no longer control chlamydial infection and start to die. Intriguingly, cellular and adaptive immune responses were impaired in the absence of C3aR. The numbers of $CD4^+$ T cells and of re-stimulated chlamydia-specific IFN-γ^+ $CD4^+$ and $CD8^+$ cells in lung-draining

Please cite this article in press as: Knittler, M.R., et al., *Chlamydia psittaci*: New insights into genomic diversity, clinical pathology, host–pathogen interaction and anti-bacterial immunity. Int. J. Med. Microbiol. (2014), http://dx.doi.org/10.1016/j.ijmm.2014.06.010

UNCORRECTED PROOF

lymph nodes were smaller; and there was no significant increase in B cells during infection. Following the same line, C3aR$^{-/-}$ mice were largely incapable of chlamydia-specific IgM or IgG production. To address causality and mechanism of C3aR-dependent protection in regard to the diminished antibody response, a serum transfer experiment was performed. Transfer of hyperimmune serum before infection could partially reverse the lethal outcome in C3aR$^{-/-}$ mice, indicating that antibody responses participate in protection against *C. psittaci* re-infection. This is in agreement with earlier reports, which indicated antibody-mediated protection in *C. trachomatis* and *C. muridarum* reinfections, suggesting a general function of antibodies in protection against secondary chlamydial infections (Morrison and Morrison, 2005; Su et al., 1997).

In contrast, transfer of *C. psittaci*-specific IgG to C3aR$^{-/-}$ mice on day 7, i.e. during ongoing acute infection, had only a very limited effect on disease progression and did not reduce lethality at all. This indicates that, in addition to the humoral response, other C3a-dependent functions must be critical for recovery and the much better survival rate of WT mice in our model of primary infection. CD4$^+$ and CD8$^+$ T cells each are sufficient for effective defense against *C. pneumoniae*, with IFN-γ as key player (Rothfuchs et al., 2004), suggesting a similar role of these T cell subsets in the defense against *C. psittaci* (see also Section “The cellular immune response: New insights into MHC I-mediated presentation of chlamydial antigens”). Thus, impaired *C. psittaci*-specific CD4$^+$ and cytotoxic CD8$^+$ T cell responses in infected C3aR$^{-/-}$ mice are presently the most likely causal link and explanation for the C3a-dependent protection in primary infection. Regulatory T cells, which suppress functions of both T cells and DCs and which can be downregulated by C3a (and C5a), might also participate in the immune dysfunction observed in the absence of C3aR in our model (for an illustration of the assumed C3aR-dependent regulation of the adaptive immune response against *C. psittaci* and protection, see Fig. 6B).

The studies described in this section showed that complement, and C3a and its receptor in particular, are critical for defense against *C. psittaci* in murine lung infection. However, it appears that the classical effector functions of the complement system are not important in this defense. Intriguingly, it seems as though C3a acts less as a weak pro-inflammatory mediator, but more as a powerful modulator of the adaptive cellular and humoral reactions against *C. psittaci*. Enhancement of specific B and T cell responses upon infection with an intracellular bacterium were identified as hitherto unknown features of complement and C3aR. These new functions might be of general immunological importance beyond chlamydial infections.

The cellular immune response: new insights into MHC I-mediated presentation of chlamydial antigens

It has been speculated that CD8$^+$T cells primed during chlamydia infection may lyse infected cells and deprive the pathogen of its intracellular niche (Balsara and Starnbach, 2007). If lysis of infected cells occurs early enough in the chlamydial developmental cycle, when the majority of organisms are non-infectious RBs, this mechanism could dramatically slow down the spread of infection within the host. In addition to this protective role of CD8$^+$ T cells during infection, there is evidence for the importance of IFN-□ release from activated CD8$^+$ T cells in controlling chlamydia infection (Starnbach et al., 1994). The initiation of the adaptive immune response (see also Section “Local immune response: Development and resolution of pulmonary lesions after *C. psittaci* infection of calf and mouse”) requires MHC-mediated antigen presentation (Raghavan et al., 2008). In particular, DCs (see also Sections “Local immune response: Development and resolution of pulmonary lesions after *C. psittaci* infection of calf and mouse” and “The innate immune response and its interaction with the adaptive response: Chlamydial infection and the complement system”) are equipped with a specialized machinery that promotes effective display of MHC/peptide complexes, rendering them the most potent stimulators of T lymphocytes (Thacker and Janssen, 2012). DCs perform a kind of sentinel function due to their ability to process internalized antigens before migration to secondary lymphoid organs, where they stimulate different sets of T cells (Thacker and Janssen, 2012). In the classical MHC I-pathway of antigen presentation, cytosolic antigens are degraded into peptide fragments by the proteasome and transported from the cytosol into the ER lumen by peptide transporter TAP (Raghavan et al., 2008). Peptides are then loaded on newly synthesized MHC I, and these complexes are released from the ER and transported to the cell surface via the Golgi (Peaper and Cresswell, 2008). DCs, however, have developed the ability to efficiently present peptides derived from endo- as well as exogenous antigens on MHC I via a process called cross-presentation (Neefjes and Sadaka, 2012). DCs are among the first professional antigen presenting cells (APCs) that are encountered by chlamydiae during infection (Gervassi et al., 2004) (see also Section “Local immune response: Development and resolution of pulmonary lesions after *C. psittaci* infection of calf and mouse”). It is thought that T cells primed by infected DCs play an important role in the effective immune response against chlamydia infection (Ojcius et al., 1998). The strategies used by the mammalian adaptive immune system (e.g. human and murine) are also mirrored in other jawed vertebrates, such as birds, so that, generally speaking, the avian innate and adaptive immune systems strongly resemble those of mammals (Foster and Berndt, 2013; Rolstad and Fossum, 1990; Sharma, 1997). Notably, pathogen-specific T and B cell responses combating chlamydial (Vanrompay et al., 1999) or other bacterial infections (Carvajal et al., 2008) and providing protection were also observed in birds. Moreover, since the avian immune system is also comprised of B, T, and NK cells, DCs, macrophages, as well as cyto/chemokines, it seems straightforward to speculate that mammals and birds use comparable immune mechanisms to elicit anti-chlamydial immunity.

In recent studies, a mouse model (see also Sections “Local immune response: Development and resolution of pulmonary lesions after *C. psittaci* infection of calf and mouse” and “The innate immune response and its interaction with the adaptive response: Chlamydial infection and the complement system”) was used to analyze cellular and molecular requirements of MHC I-cross presentation of *C. psittaci*-infected myeloid (m)DCs (Fiegl et al., 2013). These DCs display morphological as well as functional maturation, which depend mainly on TLR2 signaling and are characterized by formation of dendritic veils, decreased phagocytic activity, elevated expression of distinct activation/maturation markers, as well as secretion of mDC-specific chemo/cytokines that are associated with optimal antigen presentation to T cells and the clearance of infection (Fiegl et al., 2013). Most interestingly, infected DCs permit only a restricted developmental cycle of chlamydiae. Thus, multiple small inclusions containing predominantly RBs, also at late stages of cell infection, are a typical feature (Fiegl et al., 2013). A controlled survival of chlamydiae in infected DCs might be critical to ensure a broad antigen reservoir allowing sustained and efficient chlamydial antigen presentation. The restricted growth of chlamydiae in DCs is induced by TNF-α that interferes with the developmental cycle. The limited growth of *C. psittaci* in infected DCs is accompanied by retention of the chlamydial protease CPAF (see also Section “Host adaptation of *C. psittaci*: Escape from the avian innate immune response”) within bacterial compartments, structural disintegration of inclusions and autophagosomal degradation of released free cytosolic chlamydiae (Fiegl et al., 2013). CPAF had been postulated to be the main virulence factor of chlamydiae (Shaw et al., 2002). It was proposed that the bacterial protease

Please cite this article in press as: Knittler, M.R., et al., *Chlamydia psittaci*: New insights into genomic diversity, clinical pathology, host–pathogen interaction and anti-bacterial immunity. Int. J. Med. Microbiol. (2014), http://dx.doi.org/10.1016/j.ijmm.2014.06.010

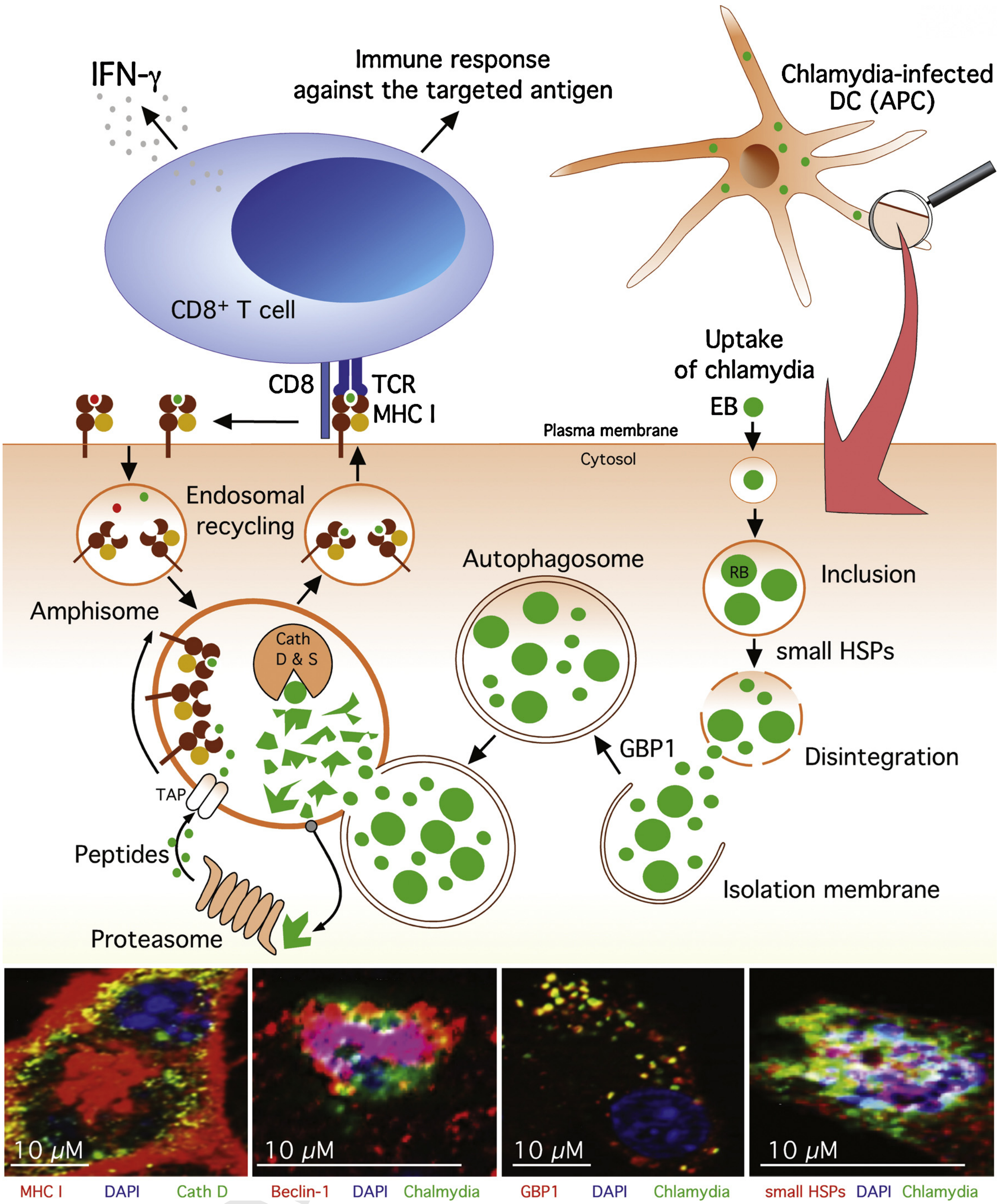

IFN-γ
Immune response
against the targeted antigen
Chlamydia-infected
DC (APC)
CD8⁺ T cell
CD8
TCR
MHC I
Uptake
of chlamydia
EB
Plasma membrane
Cytosol
Endosomal
recycling
Autophagosome
RB
Inclusion
Amphisome
Cath
D & S
small HSPs
GBP1
Disintegration
TAP
Peptides
Isolation membrane
Proteasome
10 μM
10 μM
10 μM
10 μM
MHC I DAPI Cath D
Beclin-1 DAPI Chalmydia
GBP1 DAPI Chlamydia
small HSPs DAPI Chlamydia

Fig. 7. Amphisomal cross-presentation of chlamydial antigens. Chlamydial inclusions formed in infected DCs are characterized by structural disintegration. Cytosolic released bacteria are engulfed by autophagosomes, which are able to fuse with endosomal compartments. Chlamydial antigens are preprocessed in these amphisomes and are then transferred across endo-vacuolar membranes for final processing by the proteasome. Antigenic peptides are reimported by TAP into amphisomal compartments and bound on MHC I derived from endosomal recycling of surface molecules. MHC I successfully loaded with chlamydial antigens is presented on the cell surface to CD8⁺ T cells from which they recycle back to the "loading compartment". Depicted immunofluorescence images show colocalization of chlamydia with different effectors of the host cell defense machinery including small HSPs, GBP1 as well as autophagy markers and the presence of MHC I in cathepsin D-positive compartments.

degraded various host factors relevant for chlamydial intracellular survival and that the cytosolic presence of CPAF played an essential role in maintaining the structural integrity of bacterial inclusions (Jorgensen et al., 2011). However, the findings of a recent study by Chen et al. (2012) have raised significant doubt that the identified host factors represent real substrates for CPAF in infected host cells. These new data exact a re-interpretation of the role of this bacterial enzyme during infection and fresh investigations on any CPAF functions.

By analyzing the stress response of *Chlamydia*-infected DCs, we observed an infection-dependent up-regulation of cytoprotective small HSPs (heat-shock proteins), which are known to interact with the cellular cytoskeleton. This process is accompanied by a pronounced co-localization between the HSPs and chlamydial structures. Further experiments showed that knockdown of small HSPs in infected cells resulted in an increase in the bacterial load of infected cells with intact parasitophorous vacuoles encapsulated by actin ring structures. We suggest that

Please cite this article in press as: Knittler, M.R., et al., *Chlamydia psittaci*: New insights into genomic diversity, clinical pathology, host–pathogen interaction and anti-bacterial immunity. Int. J. Med. Microbiol. (2014), http://dx.doi.org/10.1016/j.ijmm.2014.06.010

the observed effects of infection-induced HSPs are likely due to direct interventions on chlamydial cytoskeleton subversion and in turn structural disintegration of chlamydial compartments. Moreover, our studies revealed up-regulation of the host defense factor GBP1 in chlamydia-infected DCs, which targets cytosolic released chlamydiae to autophagosomes to confer cell-autonomous protection and defense against the pathogen (Fiegl et al., 2013). Autophagy is a fundamental cellular mechanism essentially involved in the degradation of intracellular structures (Huang and Brumell, 2009) and has a protective function against cell-invading microbes. Our results suggest that chlamydiae from disintegrated inclusions are efficiently guided to autophagosomal compartments termed amphisomes, which are in close proximity to the collapsed chlamydial structures (Fiegl et al., 2013).

It was also found that chlamydiae do apparently not interfere with the ability of DCs to present MHC I-bound antigens (Fiegl et al., 2013). Thus, *C. psittaci*-infected mature DCs showed increased expression, surface recycling and presentation of MHC I (Fiegl et al., 2013). The induction of MHC I depends on the presence of TNF-α (Fiegl et al., 2013), which is released by chlamydia-infected DCs in response to TLR2 signaling (Prebeck et al., 2001). Chlamydia-infected DCs are also characterized by increased presence of MHC I and TAP in endosomal compartments containing cathepsin D and S (Fiegl et al., 2013). Both hydrolases are known to function as important downstream-proteases in autolysosomal/amphisomal degradation (Boland et al., 2008; Pan et al., 2012). Inhibition of the two cathepsins in infected DCs has a dramatic influence on the MHC I-mediated stimulation of chlamydia-specific CD8$^+$ T cells suggesting important functions for the MHC I-processing of chlamydial antigens (Fiegl et al., 2013). Cathepsin-generated precursor peptides are thought to require further downstream-trimming for MHC I loading (Fonteneau et al., 2003). In accordance with this idea, we found that the proteasome-inhibition affects functional MHC I-presentation to chlamydia-specific CD8$^+$ T cells, indicating a significant contribution of the proteasome in the generation of chlamydial antigens. It appears that chlamydial antigens are first processed in amphisomal compartments, where cathepsins cleave bacterial proteins into large peptides. After cytosolic re-translocation, predigested peptides are further degraded by the proteasome and are then handled by TAP, which is also critically required for functional processing of chlamydial antigens. On the basis of the work described in this section, we propose a new cross-presentation model in which autophagy constitutes an important pathway promoting proteasome/TAP-dependent MHC I processing of chlamydial antigens (Fig. 7).

Clearance of chlamydiae depends on the ability of CD8$^+$ T cells to recognize epithelial cells, in which the bacteria predominantly replicate (Balsara and Starnbach, 2007) (see also Sections 3 and 4). Our recent studies demonstrate that chlamydia-infected epithelial cells retain their full ability to perform MHC I-mediated antigen presentation in the presence and absence of IFN-γ (Kägebein et al., 2013). One could imagine that, in chlamydia-infected epithelial cells, IFN-γ from activated T cells creates a situation reflecting the above-described scenario for infected DCs. Indeed, our own studies and those of other groups showed that, in infected epithelial cells, IFN-γ restricts chlamydial growth (Kägebein et al., 2013; Morrison, 2003) affects cytoplasmic CPAF translocation (Heuer et al., 2003), and induces autophagic degradation of chlamydiae (Al-Zeer et al., 2009).

Our recent findings on MHC I antigen presentation in chlamydia-infected DCs, epithelial cells and fibroblasts provide important novel insights for the development of such future strategies designed to support and maintain T cell-mediated immune responses against chlamydial infections.

Conclusions and outlook

Morbidity and mortality resulting from chlamydial diseases are of major global significance. The zoonotic potential of *C. psittaci* is an important and ongoing challenge for research and practice in human and veterinary medicine. The German national network on "Zoonotic chlamydiae – Models for chronic and persistent infections in man and animals" addressed distinct interconnected questions relating to chlamydial genomics and pathology, as well as host–pathogen interaction and anti-bacterial immunity. The individual studies of this coordinated research effort revealed

- the plasticity zone of the *C. psittaci* genome harboring key genomic correlates to species-specific adaptation to environmental niches, tissue tropism, virulence and pathogenicity,
- a broadened understanding of the complexity between pathological changes, pathophysiological interactions and the clinical outcome of an acute *C. psittaci* infection in the mammalian lung including the resulting systemic consequences,
- new insights into the early events of anti-chlamydial immune response and tissue regeneration in different hosts,
- significant differences and similarities between molecular factors and mechanisms of individual chlamydial pathogens, which might help to develop future strategies for disease control,
- the experimental basis to unravel the mechanisms underlying immune escape and host adaptation in the course of *C. psittaci* infection,
- a previously unknown important protective function of the complement system, particularly of C3a/C3aR, in *C. psittaci* infection of the lung,
- new insights into the role of C3a/C3aR on the development of an effective adaptive immune response against these (and most likely also other) intracellular bacteria, and
- new intracellular pathways of cell-intrinsic immunity and MHC I-mediated antigen presentation.

Although efficacious antimicrobial therapies exist, vaccination is still considered to be the best approach to reduce the prevalence of chlamydial infections. However, currently, there are no vaccines available against chlamydial infection despite the many efforts that have been made throughout the years. Our findings will not only help to develop innovative intervention and vaccination strategies against chlamydial infections, but also provide comparative data for other infectious agents. Clearly, further intensive research on chlamydial infections is required, especially with regard to treatment of persistent varieties, host-adaptation, co-infection with other intracellular pathogens and bacterial mechanisms of immune evasion.

Acknowledgments

This series of studies was conducted in the framework of the national research network "Zoonotic chlamydiae – Models of chronic and persistent infections in humans and animals", which was financially supported by the Federal Ministry of Education and Research (BMBF) of Germany under grants 01KI0720, 01KI1011D, 01Kl10011F, and 01KI1011G. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The authors thank Patrik Bavoil, Baltimore (MD), for his critical reading of the manuscript and helpful discussion.

References

Al-Zeer, M.A., Al-Younes, H.M., Braun, P.R., Zerrahn, J., Meyer, T.F., 2009. IFN-gamma-inducible Irga6 mediates host resistance against *Chlamydia trachomatis* via autophagy. PLoS ONE 4, e4588.

UNCORRECTED PROOF

Please cite this article in press as: Knittler, M.R., et al., *Chlamydia psittaci*: New insights into genomic diversity, clinical pathology, host–pathogen interaction and anti-bacterial immunity. Int. J. Med. Microbiol. (2014), http://dx.doi.org/10.1016/j.ijmm.2014.06.010

Andersen, A.A., Vanrompay, D., 2000. Avian chlamydiosis. Rev. Sci. Tech. 19, 396–404.

Balsara, Z.R., Starnbach, M.N., 2007. CD8+ T cell recognition of cells infected with *Chlamydia*. In: Bavoil, P.M., Wyrick, P.B. (Eds.), *Chlamydia*: Genomics, Pathogenesis and Implications for Control, Horizon Scientific Press, Norfolk, UK, pp. 381–412.

Bastidas, R.J., Elwell, C.A., Engel, J.N., Valdivia, R.H., 2013. Chlamydial intracellular survival strategies. Cold Spring Harb. Perspect. Med. 3, a010256.

Beatty, W.L., Morrison, R.P., Byrne, G.I., 1994. Persistent chlamydiae: from cell culture to a paradigm for chlamydial pathogenesis. Microbiol. Rev. 58, 686–699.

Beeckman, D.S., Vanrompay, D.C., 2009. Zoonotic *Chlamydophila psittaci* infections from a clinical perspective. Clin. Microbiol. Infect. 15, 11–17.

Beers, M.F., Morrisey, E.E., 2011. The three R's of lung health and disease: repair, remodeling, and regeneration. J. Clin. Invest. 121, 2065–2073.

Bode, J., Dutow, P., Sommer, K., Janik, K., Glage, S., Tummler, B., Munder, A., Laudeley, R., Sachse, K.W., Klos, A., 2012. A new role of the complement system: C3 provides protection in a mouse model of lung infection with intracellular *Chlamydia psittaci*. PLoS ONE 7, e50327.

Boland, B., Kumar, A., Lee, S., Platt, F.M., Wegiel, J., Yu, W.H., Nixon, R.A., 2008. Autophagy induction and autophagosome clearance in neurons: relationship to autophagic pathology in Alzheimer's disease. J. Neurosci. 28, 6926–6937.

Borel, N., Summersgill, J.T., Mukhopadhyay, S., Miller, R.D., Ramirez, J.A., Pospischil, A., 2008. Evidence for persistent *Chlamydia pneumoniae* infection of human coronary atheromas. Atherosclerosis 199, 154–161.

Borth, N., Litsche, K., Franke, C., Sachse, K., Saluz, H.P., Hanel, F., 2011. Functional interaction between type III-secreted protein IncA of *Chlamydophila psittaci* and human G3BP1. PLoS ONE 6, e16692.

Braukmann, M., Sachse, K., Jacobsen, I.D., Westermann, M., Menge, C., Saluz, H.P., Berndt, A., 2012. Distinct intensity of host–pathogen interactions in *Chlamydia psittaci*- and *Chlamydia abortus*-infected chicken embryos. Infect. Immun. 80, 2976–2988.

Carvajal, B.G., Methner, U., Pieper, J., Berndt, A., 2008. Effects of *Salmonella enterica* serovar Enteritidis on cellular recruitment and cytokine gene expression in caecum of vaccinated chickens. Vaccine 26, 5423–5433.

Chen, A.L., Johnson, K.A., Lee, J.K., Sutterlin, C., Tan, M., 2012. CPAF: a chlamydial protease in search of an authentic substrate. PLoS Pathog. 8, e1002842.

Collingro, A., Tischler, P., Weinmaier, T., Penz, T., Heinz, E., Brunham, R.C., Read, T.D., Bavoil, P.M., Sachse, K., Kahane, S., Friedman, M.G., Rattei, T., Myers, G.S., Horn, M., 2011. Unity in variety – the pan-genome of the Chlamydiae. Mol. Biol. Evol. 28, 3253–3270.

Dean, D., Rothschild, J., Ruettger, A., Kandel, R.P., Sachse, K., 2013. Zoonotic Chlamydiaceae species associated with trachoma, Nepal. Emerg. Infect. Dis. 19, 1948–1955.

Dehoux, P., Flores, R., Dauga, C., Zhong, G., Subtil, A., 2011. Multi-genome identification and characterization of chlamydiae-specific type III secretion substrates: the Inc proteins. BMC Genomics 12, 109.

Derre, I., Swiss, R., Agaisse, H., 2011. The lipid transfer protein CERT interacts with the *Chlamydia* inclusion protein IncD and participates to ER-*Chlamydia* inclusion membrane contact sites. PLoS Pathog. 7, e1002092.

Dutow, P., Fehlhaber, B., Bode, J., Laudeley, R., Rheinheimer, C., Glage, S., Wetsel, R.A., Pabst, O., Klos, A., 2013. The complement C3a receptor is critical in defense against *Chlamydia psittaci* in mouse lung infection and required for antibody and optimal T cell response. J. Infect. Dis..

Escalante-Ochoa, C., Ducatelle, R., Haesebrouck, F., 2000. Optimal development of *Chlamydophila psittaci* in L929 fibroblast and BGM epithelial cells requires the participation of microfilaments and microtubule-motor proteins. Microb. Pathog. 28, 321–333.

Everett, K.D., Bush, R.M., Andersen, A.A., 1999. Emended description of the order Chlamydiales, proposal of Parachlamydiaceae fam. nov. and Simkaniaceae fam. nov., each containing one monotypic genus, revised taxonomy of the family Chlamydiaceae, including a new genus and five new species, and standards for the identification of organisms. Int. J. Syst. Bacteriol. 49 (Pt 2), 415–440.

Fedorko, D.P., Clark, R.B., Nachamkin, I., Dalton, H.P., 1987. Complement dependent in vitro neutralization of *Chlamydia trachomatis* by a subspecies-specific monoclonal antibody. Med. Microbiol. Immunol. 176, 225–228.

Fiegl, D., Kagebein, D., Liebler-Tenorio, E.M., Weisser, T., Sens, M., Gutjahr, M., Knittler, M.R., 2013. Amphisomal route of MHC class I cross-presentation in bacteria-infected dendritic cells. J. Immunol. 190, 2791–2806.

Fonteneau, J.F., Kavanagh, D.G., Lirvall, M., Sanders, C., Cover, T.L., Bhardwaj, N., Larsson, M., 2003. Characterization of the MHC class I cross-presentation pathway for cell-associated antigens by human dendritic cells. Blood 102, 4448–4455.

Foster, N., Berndt, A., 2013. Immunity to *Salmonella* in farm animals and murine models of disease. In: Barrow, P.A., Methner, U. (Eds.), Salmonella in Domestic Animals. , 2nd ed, pp. 136–161.

Gaede, W., Reckling, K.F., Dresenkamp, B., Kenklies, S., Schubert, E., Noack, U., Irmscher, H.M., Ludwig, C., Hotzel, H., Sachse, K., 2008. *Chlamydophila psittaci* infections in humans during an outbreak of psittacosis from poultry in Germany. Zoonoses Public Health 55, 184–188.

Gervassi, A., Alderson, M.R., Suchland, R., Maisonneuve, J.F., Grabstein, K.H., Probst, P., 2004. Differential regulation of inflammatory cytokine secretion by human dendritic cells upon *Chlamydia trachomatis* infection. Infect. Immun. 72, 7231–7239.

Gieffers, J., van Zandbergen, G., Rupp, J., Sayk, F., Kruger, S., Ehlers, S., Solbach, W., Maass, M., 2004. Phagocytes transmit *Chlamydia pneumoniae* from the lungs to the vasculature. Eur. Respir. J. 23, 506–510.

Goellner, S., Schubert, E., Liebler-Tenorio, E., Hotzel, H., Saluz, H.P., Sachse, K., 2006. Transcriptional response patterns of *Chlamydophila psittaci* in different in vitro models of persistent infection. Infect. Immun. 74, 4801–4808.

Grieshaber, S.S., Grieshaber, N.A., Hackstadt, T., 2003. *Chlamydia trachomatis* uses host cell dynein to traffic to the microtubule-organizing center in a p50 dynamitin-independent process. J. Cell Sci. 116, 3793–3802.

Grinblat-Huse, V., Drabek, E.F., Creasy, H.H., Daugherty, S.C., Jones, K.M., Santana-Cruz, I., Tallon, L.J., Read, T.D., Hatch, T.P., Bavoil, P., Myers, G.S., 2011. Genome sequences of the zoonotic pathogens *Chlamydia psittaci* 6BC and Cal10. J. Bacteriol. 193, 4039–4040.

Harkinezhad, T., Verminnen, K., De Buyzere, M., Rietzschel, E., Bekaert, S., Vanrompay, D., 2009. Prevalence of *Chlamydophila psittaci* infections in a human population in contact with domestic and companion birds. J. Med. Microbiol. 58, 1207–1212.

Hatch, T.P., 1975. Utilization of L-cell nucleoside triphosphates by *Chlamydia psittaci* for ribonucleic acid synthesis. J. Bacteriol. 122, 393–400.

Heddema, E.R., van Hannen, E.J., Duim, B., de Jongh, B.M., Kaan, J.A., van Kessel, R., Lumeij, J.T., Visser, C.E., Vandenbroucke-Grauls, C.M., 2006. An outbreak of psittacosis due to *Chlamydophila psittaci* genotype A in a veterinary teaching hospital. J. Med. Microbiol. 55, 1571–1575.

Henning, K., Reinhold, P., Hotzel, H., Moser, I., 2008. Outbreak of a *Chlamydophila psittaci* infection in laboratory rats. Bull. Vet. Inst. Pulawy 52, 347–349.

Heuer, D., Brinkmann, V., Meyer, T.F., Szczepek, A.J., 2003. Expression and translocation of chlamydial protease during acute and persistent infection of the epithelial HEp-2 cells with *Chlamydophila* (*Chlamydia*) *pneumoniae*. Cell. Microbiol. 5, 315–322.

Heuer, D., Rejman Lipinski, A., Machuy, N., Karlas, A., Wehrens, A., Siedler, F., Brinkmann, V., Meyer, T.F., 2009. *Chlamydia* causes fragmentation of the Golgi compartment to ensure reproduction. Nature 457, 731–735.

Heymann, J., Rejman Lipinski, A., Bauer, B., Meyer, T.F., Heuer, D., 2013. *Chlamydia trachomatis* infection prevents front-rear polarity of migrating HeLa cells. Cell. Microbiol. 15, 1059–1069.

Hotzel, H., Berndt, A., Melzer, F., Sachse, K., 2004. Occurrence of Chlamydiaceae spp. in a wild boar (*Sus scrofa* L.) population in Thuringia (Germany). Vet. Microbiol. 103, 121–126.

Huang, J., Brumell, J.H., 2009. Autophagy in immunity against intracellular bacteria. Curr. Top. Microbiol. Immunol. 335, 189–215.

Jaeger, J., Liebler-Tenorio, E., Kirschvink, N., Sachse, K., Reinhold, P., 2007. A clinically silent respiratory infection with *Chlamydophila* spp. in calves is associated with airway obstruction and pulmonary inflammation. Vet. Res. 38, 711–728.

Jee, J., Degraves, F.J., Kim, T., Kaltenboeck, B., 2004. High prevalence of natural *Chlamydophila* species infection in calves. J. Clin. Microbiol. 42, 5664–5672.

Johnson, K.A., Tan, M., Sutterlin, C., 2009. Centrosome abnormalities during a *Chlamydia trachomatis* infection are caused by dysregulation of the normal duplication pathway. Cell. Microbiol. 11, 1064–1073.

Jorgensen, I., Bednar, M.M., Amin, V., Davis, B.K., Ting, J.P., McCafferty, D.G., Valdivia, R.H., 2011. The Chlamydia protease CPAF regulates host and bacterial proteins to maintain pathogen vacuole integrity and promote virulence. Cell Host Microbe 10, 21–32.

Joseph, S.J., Didelot, X., Rothschild, J., de Vries, H.J., Morre, S.A., Read, T.D., Dean, D., 2012. Population genomics of *Chlamydia trachomatis*: insights on drift, selection, recombination, and population structure. Mol. Biol. Evol. 29, 3933–3946.

Kägebein, D., Gutjahr, M., Grosse, C., Vogel, A.B., Rodel, J., Knittler, M.R., 2013. *Chlamydia*-infected epithelial cells and fibroblasts retain the ability to express surface-presented MHC class I molecules. Infect. Immun. (Epub 18 Dec 2013).

Kaleta, E.F., Taday, E.M., 2003. Avian host range of *Chlamydophila* spp. based on isolation, antigen detection and serology. Avian Pathol. 32, 435–461.

Kauffold, J., Melzer, F., Berndt, A., Hoffmann, G., Hotzel, H., Sachse, K., 2006. Chlamydiae in oviducts and uteri of repeat breeder pigs. Theriogenology 66, 1816–1823.

Kemmerling, K., Muller, U., Mielenz, M., Sauerwein, H., 2009. *Chlamydophila* species in dairy farms: polymerase chain reaction prevalence, disease association, and risk factors identified in a cross-sectional study in western Germany. J. Dairy Sci. 92, 4347–4354.

Klos, A., Wende, E., Wareham, K.J., Monk, P.N., International Union of Pharmacology, 2013. LXXXVII. Complement peptide C5a, C4a, and C3a receptors. Pharmacol. Rev. 65, 500–543.

Knowlton, A.E., Brown, H.M., Richards, T.S., Andreolas, L.A., Patel, R.K., Grieshaber, S.S., 2011. *Chlamydia trachomatis* infection causes mitotic spindle pole defects independently from its effects on centrosome amplification. Traffic 12, 854–866.

Kol, A., Bourcier, T., Lichtman, A.H., Libby, P., 1999. Chlamydial and human heat shock protein 60s activate human vascular endothelium, smooth muscle cells, and macrophages. J. Clin. Invest. 103, 571–577.

Kuo, C.-C., Stephens, R.S., Bavoil, P.M., Kaltenboeck, B., 2011. Genus *Chlamydia* Jones, Rake and Stearns 1945, 55. In: Krieg, N.R., Staley, J.T., Brown, D.R., Hedlund, B.P., Paster, B.J., Ward, N.L., Ludwig, W., Whitman, W.B. (Eds.), Bergey's Manual of Systematic Bacteriology, vol. 4, 2nd ed. Springer, Heidelberg, pp. 846–865.

Lambertz, J., (Doctoral thesis) 2011. Pathology and Pathogenesis of an Experimental Aerogenous Infection of Calves With *Chlamydophila psittaci* of Non-Avian Origin. University of Veterinary Medicine Hanover, TiHo, Hannover (in German) http://elib.tiho-hannover.de/dissertations/lambertzj_ss11.pdf

Laroucau, K., de Barbeyrac, B., Vorimore, F., Clerc, M., Bertin, C., Harkinezhad, T., Verminnen, K., Obeniche, F., Capek, I., Bebear, C., Durand, B., Zanella, G., Vanrompay, D., Garin-Bastuji, B., Sachse, K., 2009. Chlamydial infections in duck farms associated with human cases of psittacosis in France. Vet. Microbiol. 135, 82–89.

UNCORRECTED PROOF

Please cite this article in press as: Knittler, M.R., et al., *Chlamydia psittaci*: New insights into genomic diversity, clinical pathology, host–pathogen interaction and anti-bacterial immunity. Int. J. Med. Microbiol. (2014), http://dx.doi.org/10.1016/j.ijmm.2014.06.010

UNCORRECTED PROOF

Lenzko, H., Moog, U., Henning, K., Lederbach, R., Diller, R., Menge, C., Sachse, K., Sprague, L.D., 2011. High frequency of chlamydial co-infections in clinically healthy sheep flocks. BMC Vet. Res. 7, 29.

Lin, J.S., Yan, L.L., Ho, Y., Rice, P.A., 1992. Early complement components enhance neutralization of *Chlamydia trachomatis* infectivity by human sera. Infect. Immun. 60, 2547–2550.

Longbottom, D., Russell, M., Dunbar, S.M., Jones, G.E., Herring, A.J., 1998. Molecular cloning and characterization of the genes coding for the highly immunogenic cluster of 90-kilodalton envelope proteins from the *Chlamydia psittaci* subtype that causes abortion in sheep. Infect. Immun. 66, 1317–1324.

Lutter, E.I., Barger, A.C., Nair, V., Hackstadt, T., 2013. *Chlamydia trachomatis* inclusion membrane protein CT228 recruits elements of the myosin phosphatase pathway to regulate release mechanisms. Cell Rep. 3, 1921–1931.

Matsumoto, A., Manire, G.P., 1970. Electron microscopic observations on the effects of penicillin on the morphology of *Chlamydia psittaci*. J. Bacteriol. 101, 278–285.

Megran, D.W., Stiver, H.G., Bowie, W.R., 1985. Complement activation and stimulation of chemotaxis by *Chlamydia trachomatis*. Infect. Immun. 49, 670–673.

Möhle, K., (Doctoral thesis) 2011. Host–Pathogen Interactions in the Respiratory Tract of Calves After Intrapulmonary Infection With *Chlamydia psittaci*. University of Veterinary Medicine Hanover, TiHo, Hannover (in German) http://elib.tiho-hannover.de/dissertations/moehlek_ws11.pdf

Morrison, R.P., 2003. New insights into a persistent problem – chlamydial infections. J. Clin. Invest. 111, 1647–1649.

Morrison, S.G., Morrison, R.P., 2005. A predominant role for antibody in acquired immunity to chlamydial genital tract reinfection. J. Immunol. 175, 7536–7542.

Moulder, J.W., 1962. Some basic properties of the psittacosis-lymphogranuloma venereum group of agents. Structure and chemical composition of isolated particles. Ann. N. Y. Acad. Sci. 98, 92–99.

Narita, T., Wyrick, P.B., Manire, G.P., 1976. Effect of alkali on the structure of cell envelopes of *Chlamydia psittaci* elementary bodies. J. Bacteriol. 125, 300–307.

Neefjes, J., Sadaka, C., 2012. Into the intracellular logistics of cross-presentation. Front. Immunol. 3, 31.

Ojcius, D.M., Bravo de Alba, Y., Kanellopoulos, J.M., Hawkins, R.A., Kelly, K.A., Rank, R.G., Dautry-Varsat, A., 1998. Internalization of *Chlamydia* by dendritic cells and stimulation of *Chlamydia*-specific T cells. J. Immunol. 160, 1297–1303.

Q4 Ostermann, C., Linde, S., Siegling-Vlitakis, C., Reinhold, P., 2014. Evaluation of pulmonary dysfunctions and acid base imbalances induced by *Chlamydia psittaci* in a bovine model of respiratory infection. Multidiscip. Respir. Med. (in press).

Ostermann, C., Ruttger, A., Schubert, E., Schrodl, W., Sachse, K., Reinhold, P., 2013a. Infection, disease, and transmission dynamics in calves after experimental and natural challenge with a bovine *Chlamydia psittaci* isolate. PLoS ONE 8, e64066.

Ostermann, C., Schroedl, W., Schubert, E., Sachse, K., Reinhold, P., 2013b. Dose-dependent effects of *Chlamydia psittaci* infection on pulmonary gas exchange, innate immunity and acute-phase reaction in a bovine respiratory model. Vet. J. 196, 351–359.

Page, L.A., 1968. Proposal for the recognition of two species in the genus *Chlamydia* Jones, Rake and Stearns 1945. Int. J. Syst. Bacteriol. 18, 51–66.

Pan, L., Li, Y., Jia, L., Qin, Y., Qi, G., Cheng, J., Qi, Y., Li, H., Du, J., 2012. Cathepsin S deficiency results in abnormal accumulation of autophagosomes in macrophages and enhances Ang II-induced cardiac inflammation. PLoS ONE 7, e35315.

Pantchev, A., Sting, R., Bauerfeind, R., Tyczka, J., Sachse, K., 2010. Detection of all *Chlamydophila* and *Chlamydia* spp. of veterinary interest using species-specific real-time PCR assays. Comp. Immunol. Microbiol. Infect. Dis. 33, 473–484.

Peaper, D.R., Cresswell, P., 2008. Regulation of MHC class I assembly and peptide binding. Annu. Rev. Cell Dev. Biol. 24, 343–368.

Peters, J., Wilson, D.P., Myers, G., Timms, P., Bavoil, P.M., 2007. Type III secretion a la *Chlamydia*. Trends Microbiol. 15, 241–251.

Pospischil, A., Borel, N., Chowdhury, E.H., Guscetti, F., 2009. Aberrant chlamydial developmental forms in the gastrointestinal tract of pigs spontaneously and experimentally infected with *Chlamydia suis*. Vet. Microbiol. 135, 147–156.

Poudel, A., Elsasser, T.H., Rahman, Kh.S., Chowdhury, E.U., Kaltenboeck, B., 2012. Asymptomatic endemic *Chlamydia pecorum* infections reduce growth rates in calves by up to 48 percent. PLoS ONE 7, e44961.

Prebeck, S., Kirschning, C., Durr, S., da Costa, C., Donath, B., Brand, K., Redecke, V., Wagner, H., Miethke, T., 2001. Predominant role of toll-like receptor 2 versus 4 in *Chlamydia pneumoniae*-induced activation of dendritic cells. J. Immunol. 167, 3316–3323.

Raghavan, M., Del Cid, N., Rizvi, S.M., Peters, L.R., 2008. MHC class I assembly: out and about. Trends Immunol. 29, 436–443.

Read, T.D., Joseph, S.J., Didelot, X., Liang, B., Patel, L., Dean, D., 2013. Comparative analysis of *Chlamydia psittaci* genomes reveals the recent emergence of a pathogenic lineage with a broad host range. MBio, 4.

Reinhold, P., Jaeger, J., Liebler-Tenorio, E., Berndt, A., Bachmann, R., Schubert, E., Melzer, F., Elschner, M., Sachse, K., 2008. Impact of latent infections with *Chlamydophila* species in young cattle. Vet. J. 175, 202–211.

Reinhold, P., Liebler-Tenorio, E., Sattler, S., Sachse, K., 2011a. Recurrence of *Chlamydia suis* infection in pigs after short-term antimicrobial treatment. Vet. J. 187, 405–407.

Reinhold, P., Ostermann, C., Liebler-Tenorio, E., Berndt, A., Vogel, A., Lambertz, J., Rothe, M., Ruttger, A., Schubert, E., Sachse, K., 2012. A bovine model of respiratory *Chlamydia psittaci* infection: challenge dose titration. PLoS ONE 7, e30125.

Reinhold, P., Sachse, K., Kaltenboeck, B., 2011b. Chlamydiaceae in cattle: commensals, trigger organisms, or pathogens? Vet. J. 189, 257–267.

Rolstad, B., Fossum, S., 1990. Non-adaptive cellular immune responses as studied in euthymic and athymic nude rats. Spontaneous rejection of allogeneic lymphoid cell grafts by natural killer (NK) cells. Anat. Embryol. (Berl.) 181, 215–226.

Rothfuchs, A.G., Kreuger, M.R., Wigzell, H., Rottenberg, M.E., 2004. Macrophages, CD4+ or CD8+ cells are each sufficient for protection against *Chlamydia pneumoniae* infection through their ability to secrete IFN-gamma. J. Immunol. 172, 2407–2415.

Rzomp, K.A., Moorhead, A.R., Scidmore, M.A., 2006. The GTPase Rab4 interacts with *Chlamydia trachomatis* inclusion membrane protein CT229. Infect. Immun. 74, 5362–5373.

Sachse, K., Laroucau, K., Riege, K., Wehner, S., Dilcher, M., Myers, G., Weidmann, M., Vorimore, F., Vicari, N., Magnino, S., Liebler-Tenorio, E., Ruettger, A., Hufert, F.T., Bavoil, P.M., Rosselló-Móra, R., Marz, M., 2014. Evidence for the existence of two new members of the family Chlamydiaceae and proposal of *Chlamydia avium* sp. nov. and *Chlamydia gallinacea* sp. nov. Syst. Appl. Microbiol., http://dx.doi.org/10.1016/j.syapm.2013.1012.1004.

Salisch, H., Malottki, K., Ryll, M., Hinz, K.-M., 1996. Chlamydial infections of poultry and human health. World's Poult. Sci. J., 279–308.

Schofl, G., Voigt, A., Litsche, K., Sachse, K., Saluz, H.P., 2011. Complete genome sequences of four mammalian isolates of *Chlamydophila psittaci*. J. Bacteriol. 193, 4258.

Scidmore, M.A., Hackstadt, T., 2001. Mammalian 14-3-3beta associates with the *Chlamydia trachomatis* inclusion membrane via its interaction with IncG. Mol. Microbiol. 39, 1638–1650.

Seth-Smith, H.M., Harris, S.R., Rance, R., West, A.P., Severin, J.A., Ossewaarde, J.M., Cutcliffe, L.T., Skilton, R.J., Marsh, P., Parkhill, J., Clarke, I.N., Thomson, N.R., 2011. Genome sequence of the zoonotic pathogen *Chlamydophila psittaci*. J. Bacteriol. 193, 1282–1283.

Seth-Smith, H.M., Sait, M., Sachse, K., Gaede, W., Longbottom, D., Thomson, N.R., 2013. Genome sequence of *Chlamydia psittaci* strain 01DC12 originating from swine. Genome Announc., 1.

Sharma, J.M., 1997. The structure and function of the avian immune system. Acta Vet. Hung. 45, 229–238.

Shaw, A.C., Vandahl, B.B., Larsen, M.R., Roepstorff, P., Gevaert, K., Vandekerckhove, J., Christiansen, G., Birkelund, S., 2002. Characterization of a secreted *Chlamydia* protease. Cell. Microbiol. 4, 411–424.

Starnbach, M.N., Bevan, M.J., Lampe, M.F., 1994. Protective cytotoxic T lymphocytes are induced during murine infection with *Chlamydia trachomatis*. J. Immunol. 153, 5183–5189.

Stephens, R.S., 2003. The cellular paradigm of chlamydial pathogenesis. Trends Microbiol. 11, 44–51.

Stephens, R.S., Myers, G., Eppinger, M., Bavoil, P.M., 2009. Divergence without difference: phylogenetics and taxonomy of *Chlamydia* resolved. FEMS Immunol. Med. Microbiol. 55, 115–119.

Su, H., Feilzer, K., Caldwell, H.D., Morrison, R.P., 1997. *Chlamydia trachomatis* genital tract infection of antibody-deficient gene knockout mice. Infect. Immun. 65, 1993–1999.

Szeredi, L., Hotzel, H., Sachse, K., 2005. High prevalence of chlamydial (*Chlamydophila psittaci*) infection in fetal membranes of aborted equine fetuses. Vet. Res. Commun. 29 (Suppl. 1), 37–49.

Tan, C., Spitznagel, J.K., Shou, H., Hsia, R.C., Bavoil, P.M., 2006. The polymorphic membrane protein gene family of the Chlamydiaceae. In: Bavoil, P.M., Wyrick, P.B. (Eds.), *Chlamydia*: Genomics and Pathogenesis, Horizon Bioscience, Norfolk, UK, pp. 195–218.

Thacker, R.I., Janssen, E.M., 2012. Cross-presentation of cell-associated antigens by mouse splenic dendritic cell populations. Front. Immunol. 3, 41.

Theegarten, D., Sachse, K., Mentrup, B., Fey, K., Hotzel, H., Anhenn, O., 2008. *Chlamydophila* spp. infection in horses with recurrent airway obstruction: similarities to human chronic obstructive disease. Respir. Res. 9, 14.

Thomson, N.R., Yeats, C., Bell, K., Holden, M.T., Bentley, S.D., Livingstone, M., Cerdeno-Tarraga, A.M., Harris, B., Doggett, J., Ormond, D., Mungall, K., Clarke, K., Feltwell, T., Hance, Z., Sanders, M., Quail, M.A., Price, C., Barrell, B.G., Parkhill, J., Longbottom, D., 2005. The *Chlamydophila abortus* genome sequence reveals an array of variable proteins that contribute to interspecies variation. Genome Res. 15, 629–640.

Van Lent, S., Piet, J.R., Beeckman, D., van der Ende, A., Van Nieuwerburgh, F., Bavoil, P., Myers, G., Vanrompay, D., Pannekoek, Y., 2012. Full genome sequences of all nine *Chlamydia psittaci* genotype reference strains. J. Bacteriol. 194, 6930–6931.

Vanrompay, D., Cox, E., Volckaert, G., Goddeeris, B., 1999. Turkeys are protected from infection with *Chlamydia psittaci* by plasmid DNA vaccination against the major outer membrane protein. Clin. Exp. Immunol. 118, 49–55.

Vanrompay, D., Geens, T., Desplanques, A., Hoang, T.Q., De Vos, L., Van Loock, M., Huyck, E., Mirry, C., Cox, E., 2004. Immunoblotting, ELISA and culture evidence for Chlamydiaceae in sows on 258 Belgian farms. Vet. Microbiol. 99, 59–66.

Voigt, A., Schofl, G., Heidrich, A., Sachse, K., Saluz, H.P., 2011. Full-length de novo sequence of the *Chlamydophila psittaci* type strain, 6BC. J. Bacteriol. 193, 2662–2663.

Voigt, A., Schofl, G., Saluz, H.P., 2012. The *Chlamydia psittaci* genome: a comparative analysis of intracellular pathogens. PLoS ONE 7, e35097.

Vorimore, F., Hsia, R.C., Huot-Creasy, H., Bastian, S., Deruyter, L., Passet, A., Sachse, K., Bavoil, P., Myers, G., Laroucau, K., 2013. Isolation of a new *Chlamydia* species from the Feral Sacred Ibis (*Threskiornis aethiopicus*): *Chlamydia ibidis*. PLoS ONE 8, e74823.

Wolf, K., Plano, G.V., Fields, K.A., 2009. A protein secreted by the respiratory pathogen *Chlamydia pneumoniae* impairs IL-17 signalling via interaction with human Act1. Cell. Microbiol. 11, 769–779.

Wyrick, P.B., 2010. *Chlamydia trachomatis* persistence in vitro: an overview. J. Infect. Dis. 201 (Suppl. 2), S88–S95.

Please cite this article in press as: Knittler, M.R., et al., *Chlamydia psittaci*: New insights into genomic diversity, clinical pathology, host–pathogen interaction and anti-bacterial immunity. Int. J. Med. Microbiol. (2014), http://dx.doi.org/10.1016/j.ijmm.2014.06.010

Wyrick, P.B., Brownridge, E.A., 1978. Growth of *Chlamydia psittaci* in macrophages. Infect. Immun. 19, 1054–1060.

Yin, L., Kalmar, I.D., Lagae, S., Vandendriessche, S., Vanderhaeghen, W., Butaye, P., Cox, E., Vanrompay, D., 2013. Emerging *Chlamydia psittaci* infections in the chicken industry and pathology of *Chlamydia psittaci* genotype B and D strains in specific pathogen free chickens. Vet. Microbiol. 162, 740–749.

Zugel, U., Kaufmann, S.H., 1999. Role of heat shock proteins in protection from and pathogenesis of infectious diseases. Clin. Microbiol. Rev. 12, 19–39.

UNCORRECTED PROOF

Please cite this article in press as: Knittler, M.R., et al., *Chlamydia psittaci*: New insights into genomic diversity, clinical pathology, host–pathogen interaction and anti-bacterial immunity. Int. J. Med. Microbiol. (2014), http://dx.doi.org/10.1016/j.ijmm.2014.06.010

REVIEW

CNS Neuroscience & Therapeutics

Genomics and Pharmacogenomics of Dementia

Ramón Cacabelos & Rocío Martínez-Bouza

EuroEspes Biomedical Research Center, Institute for CNS Disorders and Genomic Medicine, Bergondo, Coruña, Spain and EuroEspes Chair of Biotechnology and Genomics, Camilo José Cela University, Madrid, Spain

Keywords
Alzheimer's disease; Dementia; Genetics; Genomics; Pharmacogenetics; Pharmacogenomics.

Correspondence
Prof. Dr. Ramón Cacabelos,
EuroEspes Biomedical Research Center,
Institute for CNS Disorders and Genomic
Medicine, 15165-Bergondo, Coruña, Spain.
Tel.: +34-981-780505;
Fax: +34-981-780511;
E-mail: rcacabelos@euroespes.com

doi: 10.1111/j.1755-5949.2010.00189.x

SUMMARY

Dementia is a major problem of health in developed countries, and a prototypical paradigm of chronic disability, high cost, and social-family burden. Approximately, 10–20% of direct costs in this kind of neuropathology are related to pharmacological treatment, with a moderate responder rate below 30% and questionable cost-effectiveness. Over 200 different genes have been associated with the pathogenesis of dementia. Studies on structural and functional genomics, transcriptomics, proteomics and metabolomics have revealed the paramount importance of these novel technologies for the understanding of pathogenic cascades and the prediction of therapeutic outcomes in dementia. About 10–30% of Western populations are defective in genes of the *CYP* superfamily. The most frequent *CYP2D6* variants in the Iberian peninsula are the *1/*1 (57.84%), *1/*4 (22.78%), *1×N/*1 (6.10%), *4/*4 (2.56%), and *1/*3 (2.01%) genotypes, accounting for more than 80% of the population. The frequency of extensive (EMs), intermediate (IMs), poor (PMs), and ultra-rapid metabolizers (UMs) is about 59.51%, 29,78%, 4.46%, and 6.23%, respectively, in the general population, and 57.76, 31.05%, 5.27%, and 5.90%, respectively, in AD cases. The construction of a genetic map integrating the most prevalent *CYP2D6+CYP2C19+CYP2C9* polymorphic variants in a trigenic cluster yields 82 different haplotype-like profiles, with *1*1-*1*1-*1*1 (25.70%), *1*1-*1*2-*1*2 (10.66%), *1*1-*1*1-*1*1 (10.45%), *1*4-*1*1-*1*1 (8.09%), *1*4-*1*2-*1*1 (4.91%), *1*4-*1*1-*1*2 (4.65%), and *1*1-*1*3-*1*3 (4.33%), as the most frequent genotypes. Only 26.51% of AD patients show a pure 3EM phenotype, 15.29% are 2EM1IM, 2.04% are pure 3IM, 0% are pure 3PM, and 0% are 1UM2PM. EMs and IMs are the best responders, and PMs and UMs are the worst responders to a combination therapy with cholinesterase inhibitors, neuroprotectants, and vasoactive substances. The pharmacogenetic response in AD appears to be dependent upon the networking activity of genes involved in drug metabolism and genes involved in AD pathogenesis (e.g., *APOE*). AD patients harboring the *APOE-4/4* genotypes are the worst responders to conventional antidementia drugs. To achieve a mature discipline of pharmacogenomics in CNS disorders and dementia it would be convenient to accelerate the following processes: (i) to educate physicians and the public on the use of genetic/genomic screening in daily clinical practice; (ii) to standardize genetic testing for major categories of drugs; (iii) to validate pharmacogenomic information according to drug category and pathology; (iv) to regulate ethical, social, and economic issues; and (v) to incorporate pharmacogenomic procedures both to drugs in development and drugs on the market in order to optimize therapeutics.

Introduction

Senile dementia is becoming a major problem of health in developed countries, and the primary cause of disability in the elderly. Alzheimer's disease (AD) is the most frequent form of dementia (50–70%), followed by vascular dementia (30–40%), and mixed dementia (15–20%). These prevalent forms of age-related neurodegeneration affect over 25 million people at present, and probably over 75 million people will be at risk in the next 20–25 years worldwide. The prevalence of dementia increases exponentially from approximately 1% at 60–65 years of age to over 30–35% in people older than 80 years. It is very likely that in those patients older than 75–80 years of age most cases of dementia are mixed in nature (degenerative + vascular), whereas pure AD cases are very rare after 80 years of age. The average annual cost per person with dementia ranges from €10,000 to €40,000, depending upon disease stage and country, with a lifetime cost per patient of more than €150,000. In some countries, approximately 80% of the global costs of dementia (direct + indirect costs) are assumed by the patients and/or their families. About 10–20% of the costs in dementia are attributed to pharmacological treatment, including antidementia drugs, psychotropics (antidepressants, neuroleptics, and anxiolytics), and other drugs currently prescribed in the elderly (antiparkinsonians, anticonvulsants, vasoactive compounds, antiinflammatory drugs, etc.). In addition, during the past 20 years over 300 drugs have been partially or totally developed for AD, with the subsequent costs for the pharmaceutical industry, and only 5 drugs with moderate-to-poor efficacy and questionable cost-effectiveness have been approved in developed countries [1–3].

Genomics of Dementia

The genetic defects identified in AD can be classified into three main categories:

(a) Mendelian or mutational defects in genes directly linked to AD, including (i) 32 mutations in the amyloid beta (Aβ) (ABP) precursor protein (*APP*) gene (21q21)(AD1); (ii) 165 mutations in the presenilin 1 (*PSEN1*) gene (14q24.3)(AD3); and (iii) 12 mutations in the presenilin 2 (*PSEN2*) gene (1q31–q42) (AD4) [4–6]. PSEN1 and PSEN2 are important determinants of γ-secretase activity responsible for proteolytic cleavage of APP and NOTCH receptor proteins. Mendelian mutations are very rare in AD (1:1000). Mutations in exons 16 and 17 of the *APP* gene appear with a frequency of 0.30% and 0.78%, respectively, in AD patients. Similarly, *PSEN1*, *PSEN2*, and microtubule-associated protein Tau (*MAPT*)(17q21.1) mutations are present in less than 2% of the cases. Mutations in these genes confer specific phenotypic profiles to patients with dementia: amyloidogeneic pathology associated with *APP*, *PSEN1* and *PSEN2* mutations; and tauopathy associated with *MATP* mutations, representing the two major pathogenic hypotheses for AD [4,7–9].

(b) Multiple polymorphic risk variants characterized in over 200 different genes can increase neuronal vulnerability to premature death(Table 1) [4]. Among these susceptibility genes, the apolipoprotein E (*APOE*) gene (19q13.2)(AD2) is the most prevalent as a risk factor for AD, especially in those subjects harboring the *APOE-4* allele, whereas carriers of the *APOE-2* allele might be protected against dementia [4]. APOE-related pathogenic mechanisms are also associated with brain aging and with the neuropathological hallmarks of AD. Other genes of this category are included in Table 1.

One of the newest members of the AD-gene family is *SORL1*, a gene which encodes a mosaic protein with a domain structure which suggests it is a member of both the vacuolar protein sorting-10 (Vps10) domain-containing receptor family and the low density lipoprotein receptor (LDLR). Inherited variants of the SORL1 neuronal sorting receptor are associated with late-onset AD. Polymorphisms in two different clusters of intronic sequences within the *SORL1* gene may regulate tissue-specific expression of SORL1, which directs trafficking of APP into recycling pathways. When SORL1 is underexpressed, APP is sorted into Aβ-generating compartments leading to amyloid accumulation in neuronal tissues [10]. As with many other potential AD-related genes, the association of *SORL1* with AD [10,11] could not be replicated in other studies [12]. Another interesting gene is *DHCR24* (3β-hydroxysterol-δ-24-reductase) or *Seladin-1*, a key element in the cholesterologenic pathway in which the DHCR24 enzyme catalyzes the transformation of desmosterol into cholesterol [13,14]. *Seladin-1* was originally identified as a gene whose expression was downregulated in the AD brain, demonstrating a neuroprotective effect against neurodegeneration. Recent studies indicate that *Seladin-1/DHCR24* is an LXR (liver X nuclear hormone receptor) target gene potentially involved in the regulation of lipid raft formation [13]. Another gene, with potential therapeutic interest as a tau kinase, might be the *GSK3* gene. Analysis of the promoter and all 12 exons revealed that an intronic polymorphism (IVS2-68G>A) occurred at over twice the frequency among patients with frontotemporal dementia (10.8%) and patients with AD (14.6%) than in aged healthy subjects (4.1%). This is the first evidence that a gene known to be involved in tau phosphorylation is associated with risk for primary neurodegenerative dementias [15]. Promoter polymorphisms modulating *HSPA5* expression might also increase susceptibility to AD. Endoplasmic reticulum chaperone heat shock 70 kDa protein 5 (HSPA5/GRP78) is known to be involved in APP metabolism and neuronal death in AD. Of the three major polymorphisms (–415G/A (rs391957), –370C/T (rs17840761), and –180del/G (rs3216733)), the HSPA5-415G/A and –180del/G variants showed significant differences between AD cases and controls. Subjects harboring the –415AA/–180GG genotype or the –415A/–180G allele might be less susceptible to develop AD [16]. The rs5952C and rs1568566T alleles of the *APOD* rs5952T/C and rs1568566C/T variants increase the risk for AD, whereas the rs5952T-rs1568566C haplotype reduces it [17]. ApoD is a lipoprotein-associated glycoprotein which is increased in the hippocampus and CSF of AD patients [17]. *CALHM1* encodes a multipass transmembrane glycoprotein that controls cytosolic Ca^{2+} concentrations and Aβ levels. The *CALHM1* P86L polymorphism (rs2986017) has been associated with AD [18], but this association could not be replicated in other studies.

Harold et al. [19] undertook a two-stage genome-wide association study (GWAS) of AD involving over 16,000 individuals, and found association with SNPs at two loci not previously associated with the disease, at the *CLU* (Clusterine, APOJ) gene (rs11136000) and 5′ to the *PICALM* gene (rs3851179). In another GWAS with patients from France, Belgium, Finland, Italy and Spain, Lambert et al. [20] found association with *CLU* and with the *CR1* gene, encoding the complement component (3b/4b) receptor 1, on chromosome 1 (rs6656401).

(c) Diverse mutations located in mitochondrial DNA (mtDNA) through heteroplasmic transmission can influence aging and oxidative stress conditions, conferring phenotypic heterogeneity [4,21].

Although *APP* and *PSEN* mutations are considered causative factors for AD, the total number of mutations identified in the

Table 1. Selected human genes investigated as potential candidate genes associated with dementia and age-related neurodegenerative disorders [4–6,25].

Locus	Symbol	Aliases	Title	OMIM
1p21.3-p13.1	*SORT1*	*Gp95, NT3*	Sortilin	602458
1p31.3	*TM2D1*	*BBP*	TM2 domain containing 1	610080
1p32	*ERI3*	*PINT1; PRNPIP; MGC2683; FLJ22943*	ERI1 exoribonuclease family member 3	609917
1p32.3	*ZFYVE9*	*MADHIP, NSP, SARA, SMADIP*	Zinc finger, FYVE domain containing 9	603755
1p33-p31.1	*DHCR24*	*KIAA0018, Nbla03646, SELADIN1, seladin-1*	24-dehydrocholesterol reductase	606418
1p34	*LRP8*	*APOER2, HSZ75190, MCI1*	Low density lipoprotein receptor-related protein 8, apolipoprotein e receptor	602600
1p36.3	*MTHFR*		5,10-methylenetetrahydrofolate reductase (NADPH)	607093
1q21	*S100A1*	*S100, S100-alpha, S100A*	S100 calcium-binding protein A1	176940
1q21.2-q21.3	*LMNA*	*RP11-54H19.1, CDCD1, CDDC, CMD1A, CMT2B1, EMD2, FPL, FPLD, HGPS, IDC, LDP1, LFP, LGMD1B, LMN1, LMNC, PRO1*	Lamin A/C	150330
1q21.3	*CHRNB2*	*EFNL3, nAChRB2*	Cholinergic receptor, nicotinic, beta 2 (neuronal)	118507
1q21-q23	*APCS*	*MGC88159, PTX2, SAP*	Amyloid P component, serum	104770
1q22-q23	*NCSTN*	*RP11-517F10.1, APH2, KIAA0253*	Nicastrin	605254
1q25	*SOAT1*	*RP11-215I23.1, ACACT, ACAT, ACAT1, RP11-215I23.2, SOAT, STAT*	Sterol O-acyltransferase 1	102642
1q31-q42	*AD4*	*AD3L, AD4, PS2, STM2*	Presenilin 2 (alzheimer disease 4)	600759
1q32	*CR1*	*C3BR, C4BR, CD35, KN*	Complement component (3b/4b) receptor 1 (Knops blood group)	120620
1p36.13-q31.3	*APH1A*	*RP4-790G17.3, 6530402N02Rik, APH-1, APH-1A, CGI-78*	Anterior pharynx defective 1 homolog A (C. elegans)	607629
1q42-q43	*AGT*	*ANHU, FLJ92595, FLJ97926, SERPINA8*	Angiotensinogen (serpin peptidase inhibitor, clade A, member 8)	106150
2p16.3	*RTN4*	*ASY, NI220/250, NOGO, NOGO-A, NOGOC, NSP, NSP-CL, Nbla00271, Nbla10545, Nogo-B, Nogo-C, RTN-X, RTN4-A, RTN4-B1, RTN4-B2, RTN4-C*	Reticulon 4	604475
2p25	*ADAM17*	*ADAM18, CD156B, CSVP, MGC71942, TACE*	ADAM metallopeptidase domain 17	603639
2q14	*IL1A*	*IL-1A, IL1, IL1-ALPHA, IL1F1*	Interleukin-1-Alpha	147760
2q21.1	*KCNIP3*	*CSEN, DREAM, KCHIP3, MGC18289*	Kv channel interacting protein 3, calsenilin	604662
2q21.2	*LRP1B*	*LRP-DIT, LRPDIT*	Low density lipoprotein-related protein 1B (deleted in tumors)	608766
2q34	*CREB1*	*CREB, MGC9284*	cAMP responsive element binding protein 1	123810
3q26.1-q26.2	*BCHE*	*CHE1, E1*	Butyrylcholinesterase	177400
3q32.3-q34	*CREB1*		cAMP response element-binding protein	123810
3q26.2-qter		*APOD*	Apolipoprotein D	107740
4p14-p13	*APBB2*	*DKFZp434E033, FE65L, FE65L1, MGC35575*	Amyloid beta (A4) precursor protein-binding, family B, member 2	602710
5q15	*CAST*	*BS-17, MGC9402*	Calpastatin	114090
5q31	*APBB3*	*FE65L2, MGC150555, MGC87674, SRA*	Amyloid beta (A4) precursor protein-binding, family B, member 3	602711
5q35.3	*DBN1*	*D0S117E, DKFZp434D064*	drebrin 1	126660
6p21.3	*AGER*	*DAMA-358M23.4, MGC22357, RAGE*	Advanced glycosylation end product-specific receptor	600214
6p21.3	*HFE*	*HFE1, HH, HLA-H, MGC103790, MVCD7, dJ221C16.10.1*	Hemochromatosis	235200
6p21.3	*TNF*	*DADB-70P7.1, DIF, TNF-alpha, TNFA, TNFSF2*	Tumor necrosis factor (TNF superfamily, member 2)	191160
7p21	*IL6*	*BSF2, HGF, HSF, IFNB2, IL-6*	Interleukin 6 (interferon, beta 2)	147620

Table 1. Continued

Locus	Symbol	Aliases	Title	OMIM
7q36	*AD10*		Alzheimer disease-10	609636
7q36	*NOS3*	*ECNOS, eNOS*	Nitric oxide synthase 3 (endothelial cell)	163729
7q36	*PAXIP1*	*CAGF28, CAGF29, FLJ41049, PACIP1, PAXIP1L, PTIP, TNRC2*	PAX interacting (with transcription-activation domain) protein 1	608254
8p21-p12	*CLU*	*AAG4, APOJ, CLI, KUB1, MGC24903, SGP-2, SGP2, SP-40, TRPM-2, TRPM2*	Clusterin	185430
8p22	*CTSB*	*APPS, CPSB*	Cathepsin B	116810
9q13-q21.1	*APBA1*	*D9S411E, MINT1, X11, X11A, X11ALPHA*	Amyloid beta (A4) precursor protein-binding, family A, member 1	602414
9q34.1	*DAPK1*	*DAPK, DKFZp781I035*	Death-associated protein kinase 1	600831
9q33-q34.1	*HSPA5*	*BIP, FLJ26106, GRP78, MIF2*	Heat shock 70kDa protein 5 (glucose-regulated protein, 78kDa)	138120
10p13	*AD7*		Alzheimer disease 7	606187
10q	*AD6*		Alzheimer disease-6	605526
10q21	*TFAM*	*MtTF1, TCF6, TCF6L1, TCF6L2, TCF6L3, mtTFA*	Transcription factor A, mitochondrial	600438
10q23	*CH25H*	*C25H*	Cholesterol 25-hydroxylase	604551
10q23-q25	*IDE*	*RP11-366I13.1, FLJ35968, INSULYSIN*	Insulin-degrading enzyme	146680
10q23-q25	*SORCS1*	*RP11-446H13.1, FLJ41758, FLJ43475, FLJ44957*	Sortilin-related VPS10 domain containing receptor 1	606283
10q24	*PLAU*	*ATF, UPA, URK, u-PA*	Plasminogen activator, urokinase	191840
10q24.33	*CALHM1*	*FAM26C, MGC39514, MGC39617*	Calcium homeostasis modulator 1	612234
11p13	*BDNF*	*MGC34632*	Brain-derived neurotrophic factor	113505
11p15	*APBB1*	*FE65, MGC:9072, RIR*	Amyloid beta (A4) precursor protein-binding, family B, member 1 (Fe65)	602709
11p15.1	*SAA1*	*MGC111216, PIG4, SAA, TP53I4*	Serum amyloid A1	104750
11p15.5	*CTSD*	*CLN10, CPSD, MGC2311*	Cathepsin D	116840
11q14	*PICALM*	*CALM, CLTH, LAP*	Phosphatidylinositol binding clathrin assembly protein	603025
11q23.2-q24.2	*SORL1*	*C11orf32, FLJ21930, FLJ39258, LR11, LRP9, SORLA, SorLA-1, gp250*	sortilin-related receptor, L(DLR class) A repeats-containing	602005
11q23.2-q23.3	*BACE1*	*ASP2, BACE, FLJ90568, HSPC104, KIAA1149*	Beta-site APP-cleaving enzyme 1	604252
11q24	*APLP2*	*APPH, APPL2, CDEBP*	Amyloid beta (A4) precursor-like protein 2	104776
12p11.23-q13.12	*AD5*		Alzheimer disease 5	602096
12p12.3-p12.1	*IAPP*	*AMYLIN, DAP, IAP*	Islet amyloid polypeptide	147940
12p13.3-p12.3	*A2M*	*CPAMD5, DKFZp779B086, FWP007, S863-7*	Alpha-2-macroglobulin	103950
12q13-q14	*LRP1*	*A2MR, APOER, APR, CD91, FLJ16451, IGFBP3R, LRP, MGC88725, TGFBR5*	Low density lipoprotein-related protein 1 (alpha-2-macroglobulin receptor)	107770
14q24.3	*FOS*	*AP-1, C-FOS*	FBJ murine osteosarcoma viral oncogene homolog	164810
14q24.3	*PSEN1*	*AD3, FAD, PS1, S182*	presenilin-1	104311
14q32.1	*SERPINA3*	*AACT, ACT, GIG24, GIG25, MGC88254*	Serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 3	107280
14q32.1	*CYP46A1*	*CP46, CYP46*	Cytochrome P450, family 46, subfamily A, polypeptide 1	604087
15q22.2	*APH1B*	*APH-1B, DKFZp564D0372, FLJ33115, PRO1328, PSFL, TAAV688*	Anterior pharynx defective 1 homolog B (C. elegans)	607630
15q11-q12	*APBA2*	*D15S1518E, HsT16821, LIN-10, MGC99508, MGC:14091, MINT2, X11L*	Amyloid beta (A4) precursor protein-binding, family A, member 2	602712
16q21	*CETP*	*HDLCQ10*	Cholesteryl ester transfer protein, plasma	118470

Table 1. Continued

Locus	Symbol	Aliases	Title	OMIM
16q22	*NAE1*	*A-116A10.1, APPBP1, HPP1, ula-1*	NEDD8 activating enzyme E1 subunit 1	603385
17p13	*MYH13*	*MyHC-eo*	Myosin, heavy chain 13, skeletal muscle	603487
17p13.1	*TNK1*	*MGC46193*	Tyrosine kinase, non-receptor, 1	608076
17q11.2	*BLMH*	*BH, BMH*	Bleomycin hydrolase	602403
17q21.1	*MAPT*	*DDPAC, FLJ31424, FTDP-17, MAPTL, MGC138549, MSTD, MTBT1, MTBT2, PPND, TAU*	Microtubule-associated protein tau	157140
17q21.1	*STH*	*MAPTIT, MGC163191, MGC163193*	Saitohin	607067
17q21-q22	*GPSC*		Gliosis, familial progressive subcortical	221820
17q21-q23	*APPBP2*	*HS.84084, KIAA0228, PAT1*	Amyloid beta precursor protein (cytoplasmic tail) binding protein 2	605324
17q23.3	*ACE*	*ACE1, CD143, DCP, DCP1, MGC26566, MVCD3*	Angiotensin I converting enzyme (peptidyl-dipeptidase A) 1	106180
17q23.1	*MPO*		Myeloperoxidase	606989
17q24.3	*BPTF*	*FAC1, FALZ, NURF301*	Bromodomain PHD finger transcription factor	601819
18q12.1	*TTR*	*HsT2651, PALB, TBPA*	Transthyretin	176300
19p13.2-p13.1	*NOTCH3*	*CADASIL, CASIL*	Notch homolog 3 (Drosophila)	600276
19p13.2	*AD9*		Alzheimer disease 9	608907
19p13.3	*APBA3*	*MGC:15815, X11L2, mint3*	Amyloid beta (A4) precursor protein-binding, family A, member 3	604262
19p13.3-p13.2	*ICAM*	*BB2, CD54, P3.58*	Intercellular adhesion molecule 1	147840
19q13.1	*APLP1*	*APLP*	Amyloid beta (A4) precursor-like protein 1	104775
19q13.12	*PEN2*	*MDS033, MSTP064, PEN-2, PEN2*	Presenilin enhancer 2 homolog (C. elegans)	607632
19q13.2	*APOE*	*AD2, LDLCQ5, LPG, MGC1571*	Apolipoprotein E	107741
19q13.2	*APOC1*		Apolipoprotein C-I	107710
20p	*AD8*		Alzheimer disease-8	607116
20p11.21	*CST3*	*ARMD11, MGC117328*	Cystatin C	604312
20p13	*PRNP*	*ASCR, CD230, CJD, GSS, MGC26679, PRIP, PrP, PrP27-30, PrP33-35C, PrPc, prion*	Prion protein	176640
20q13.31	*PCK1*	*MGC22652, PEPCK-C, PEPCK1, PEPCKC*	Phosphoenolpyruvate carboxykinase 1 (soluble)	261680
21q21.3	*APP*	*AAA, ABETA, ABPP, AD1, APPI, CTFgamma, CVAP, PN2*	Amyloid beta (A4) precursor protein	104760
21q22.3	*BACE2*	*AEPLC, ALP56, ASP1, ASP21, BAE2, CDA13, CEAP1, DRAP*	Beta-site APP-cleaving enzyme 2	605668
22q11.21	*RTN4R*	*NGR, NOGOR*	Reticulon 4 receptor	605566
	HN		Humanin	606120

APP, *PSEN1*, and *PSEN2* genes account for less than 3% of the cases with AD, clearly indicating that neurodegeneration associated with AD pathogenesis cannot be exclusively attributed to APP/PSEN-related cascades (amyloid hypothesis). Alterations in the ubiquitin–proteasome system and biochemical disarray in the chaperone machinery are alternative and/or complementary pathogenic events potentially leading to defects in protein synthesis, folding, and degradation with subsequent conformational changes, aggregation, and accumulation in cytotoxic deposits [4,22]. A more plausible explanation would seem to be that multiple susceptibility SNPs with a very subtle genetic variation cooperatively contribute, in concert with environmental factors and concomitant CNS vulnerability, to premature neurodegeneration in dementia.

It is also likely that defective functions of genes associated with longevity may influence premature neuronal survival, since neurons are potential pacemakers defining life span in mammals [4,23]. All these genetic factors may interact in genetic networks which are still unknown, leading to a cascade of pathogenic events characterized by abnormal protein processing and misfolding with a subsequent accumulation of abnormal proteins (conformational changes), ubiquitin–proteasome system dysfunction, excitotoxic reactions, oxidative and nitrosative stress, mitochondrial injury, synaptic failure, altered metal homeostasis, dysfunction of axonal and dendritic transport, and chaperone misoperation [4,22–31]. These pathogenic events may exert an additive effect, converging in final pathways leading to premature neuronal death. Some of these mechanisms are common to several neurodegenerative disorders which differ depending upon the gene(s) affected and the involvement of specific genetic networks, together with cerebrovascular factors, epigenetic factors (DNA methylation) and environmental conditions (nutrition, toxicity, social factors, etc.) [4,23,26–32]. The higher the number of genes involved in AD pathogenesis, the earlier the onset of the disease, the faster its clinical course, and the poorer its therapeutic outcome [4,23,26–31].

Functional Genomics

Functional genomics studies have demonstrated the influence of many genes on AD pathogenesis and phenotype expression. The study of genotype–phenotype correlations is essential for the evaluation of the actual impact of specific polymorphic variants of a particular gene on the clinical manifestation of the disease and/or biological markers reflecting the disease condition or different biological states of the individual. It has been demonstrated that mutations in the *APP*, *PSEN1*, *PSEN2*, and *MAPT* genes give rise to well-characterized differential neuropathological and clinical phenotypes of dementia [4–6]. Transgenic animals also reproduce to some extent the neuropathological hallmarks of AD in a sequential manner. The triple transgenic mouse model of AD (3×Tg-AD) harbors 3 AD-related loci: human PS1M146V, human APPswe, and human MAPTP301L. These animals develop both amyloid plaques and NFT-like pathology in a progressive and age-dependent manner in hippocampus, amygdala, and cerebral cortex, the main foci of human AD neuropathology. The evolution of AD-related transgene expression, amyloid deposition, tau phosphorylation, astrogliosis, and microglia activation throughout the hippocampus, entorhinal cortex, primary motor cortex, and amygdala over a 26-month period has been immunohistochemically documented. Intracellular Aβ accumulation is the earliest of AD-related pathologies to be detectable, followed temporally by phospho-tau, extracellular Aβ, and finally paired helical filament and NFT pathology [33]. In the same model, a decrease in neurogenesis directly associated with the presence of amyloid plaques and an increase in the number of Aβ containing neurons in the hippocampus has been demonstrated [34].

Different *APOE* genotypes also confer specific phenotypic profiles to AD patients. Some of these profiles may add risk or benefit when the patients are treated with conventional drugs, and in many instances the clinical phenotype demands the administration of additional drugs, which increase the complexity of therapeutic protocols. From studies designed to define *APOE*-related AD phenotypes [4,22–31,35–41], several confirmed conclusions can be drawn: (i) the age-at-onset is 5–10 years earlier in approximately 80% of AD cases harboring the *APOE-4/4* genotype; (ii) the serum levels of ApoE are lowest in *APOE-4/4*, intermediate in *APOE-3/3* and *APOE-3/4*, and highest in *APOE-2/3* and *APOE-2/4;* (iii) serum cholesterol levels are higher in APOE-4/4 than in the other genotypes; (iv) HDL-cholesterol levels tend to be lower in *APOE-3* homozygotes than in *APOE-4* allele carriers; (v) LDL-cholesterol levels are systematically higher in *APOE-4/4* than in any other genotype; (vi) triglyceride levels are significantly lower in *APOE-4/4*; (vii) nitric oxide levels are slightly lower in *APOE-4/4*; (viii) serum Aβ levels do not differ between *APOE-4/4* and the other most frequent genotypes (*APOE-3/3*, *APOE-3/4*); (ix) blood histamine levels are dramatically reduced in *APOE-4/4* as compared with the other genotypes; (x) brain atrophy is markedly increased in *APOE-4/4>APOE-3/4>APOE-3/3*; (xi) brain mapping activity shows a significant increase in slow wave activity in *APOE-4/4* from early stages of the disease; (xii) brain hemodynamics, as reflected by reduced brain blood flow velocity and increased pulsatility and resistance indices, is significantly worse in *APOE-4/4* (and in *APOE-4* carriers, in general, as compared with *APOE-3* carriers); (xiii) lymphocyte apoptosis is markedly enhanced in *APOE-4* carriers; (xiv) cognitive deterioration is faster in *APOE-4/4* patients than in carriers of any other *APOE* genotype; (xv) occasionally, in approximately 3–8% of the AD cases, the presence of some dementia-related metabolic dysfunctions (e.g., iron, folic acid, vitamin B_{12} deficiencies) accumulate more in *APOE-4* carriers than in *APOE-3* carriers; (xvi) some behavioral disturbances (bizarre behaviors, psychotic symptoms), alterations in circadian rhythm patterns (e.g., sleep disorders), and mood disorders (anxiety, depression) are slightly more frequent in *APOE-4* carriers; (xvii) aortic and systemic atherosclerosis is also more frequent in *APOE-4* carriers; (xviii) liver metabolism and transaminase activity also differ in *APOE-4/4* with respect to other genotypes; (xix) blood pressure (hypertension) and other cardiovascular risk factors also accumulate in *APOE-4*; and (xx) *APOE-4/4* are the poorest responders to conventional drugs. These 20 major phenotypic features clearly illustrate the biological disadvantage of *APOE-4* homozygotes and the potential consequences that these patients may experience when they receive pharmacological treatment [4,9,22–31,35–43].

Therapeutic Strategies in Dementia

Modern therapeutic strategies in AD are addressed to interfering with the main pathogenic mechanisms potentially involved in AD. Major pathogenic events (drug targets) and their respective therapeutic alternatives include the following: genetic defects, β-amyloid deposition, tau-related pathology, apoptosis, neurotransmitter deficits, neurotrophic deficits, neuronal loss, neuroinflammation, oxidative stress, calcium dysmetabolism, neuronal hypometabolism, lipid metabolism dysfunction, cerebrovascular dysfunction, neuronal dysfunction associated with nutritional and/or metabolic deficits, and a miscellany of pathogenic mechanisms potentially manageable with diverse classes of chemicals or biopharmaceuticals [4,22–31,35,36,41]. Since the early 1980s, the neuropharmacology of AD was dominated by the acetylcholinesterase inhibitors, represented by tacrine, donepezil, rivastigmine, and galantamine [2,3,44]. Memantine, a partial NMDA antagonist, was introduced in the 2000s for the treatment of severe dementia [45]; and the first clinical trials with immunotherapy, to reduce amyloid burden in senile plaques, were withdrawn due to severe ADRs [46]. During the past few years no relevant drug candidates have been postulated for the treatment of AD, despite the initial promises of β- and γ-secretase inhibitors [22].

Pharmacogenomics

Pharmacogenomics relates to the application of genomic technologies, such as genotyping, gene sequencing, gene expression, genetic epidemiology, transcriptomics, proteomics, metabolomics, and bioinformatics, to drugs in clinical development and on the market, applying the large-scale systematic approaches of genomics to speed up the discovery of drug response markers, whether they act at the level of drug target, drug metabolism, or disease pathways [23,25,31,35,47].

The potential implications of pharmacogenomics in clinical trials and molecular therapeutics is that a particular disease could be treated according to genomic and biological markers, selecting medications and diseases which are optimized for individual patients or clusters of patients with a similar genomic profile.

The pharmacogenomic outcome depends upon many different determinant factors including (i) genomic profile, (ii) disease phenotype, (iii) concomitant pathology, (iv) genotype–phenotype correlations, (v) nutritional conditions, (vi) age and gender, (vii) pharmacological profile of the drugs, (viii) drug–drug interactions, (ix) gene expression profile, (x) transcriptomic cascade, (xi) proteomic profile, and (xii) metabolomic networking. The dissection and further integration of all these factors is of paramount importance for the assessment of the pharmacogenomic outcome in terms of safety and efficacy. Pharmacogenomic approaches based on genomewide sets of SNPs associated with drug response are now feasible and may offer the potential to personalize therapeutics [25].

Drug metabolism includes phase I reactions (i.e., oxidation, reduction, hydrolysis) and phase II conjugation reactions (i.e., acetylation, glucuronidation, sulphation, and methylation). The principal enzymes with polymorphic variants involved in phase I reactions are the following: CYP3A4/5/7, CYP2E1, CYP2D6, CYP2C19, CYP2C9, CYP2C8, CYP2B6, CYP2A6, CYP1B1, CYP1A1/2, epoxide hydrolase, esterases, NQO1 (NADPH-quinone oxidoreductase), DPD (dihydropyrimidine dehydrogenase), ADH (alcohol dehydrogenase), and ALDH (aldehyde dehydrogenase). Major enzymes involved in phase II reactions include the following: UGTs (uridine 5′-triphosphate glucuronosyl transferases), TPMT (thiopurine methyltransferase), COMT (catechol-O-methyltransferase), HMT (histamine methyl-transferase), STs (sulfotransferases), GST-A (glutathion S-transferase A), GST-P, GST-T, GST-M, NAT2 (N-acetyl transferase), NAT1, and others [22–25].

CYPs in Dementia

In dementia, *CYP* genomics is a very important issue since in practice over 90% of patients with dementia are daily consumers of psychotropics. Furthermore, some acetylcholinesterase inhibitors (the most prescribed antidementia drugs worldwide) are metabolized via CYP enzymes [24,25,39]. CYP2D6, CYP2C19, CYP2C9, and CYP3A4/5 deserve special consideration.

The CYP2D6 enzyme, encoded by a gene that maps on 22q13.1–13.2, catalyzes the oxidative metabolism of over 100 clinically important and commonly prescribed drugs such as cholinesterase inhibitors, antidepressants, neuroleptics, opioids, some β-blockers, class I antiarrhythmics, analgesics, and many other drug categories [48], acting as substrates, inhibitors or inducers with which many other drugs may potentially interact, this leading to the outcome of ADRs. The CYP2D6 locus is highly polymorphic, with over 100 different *CYP2D6* alleles identified in the general population showing deficient (PM), normal (EM), intermediate (IM), or increased enzymatic activity (UM) [49,50]. Most individuals (>80%) are EMs; however, remarkable interethnic differences exist in the frequency of the PM and UM phenotypes among different societies all over the world [23,51,52]. On average, approximately 6.28% of the world population belongs to the PM category. Europeans (7.86%), Polynesians (7.27%), and Africans (6.73%) exhibit the highest rate of PMs, whereas Orientals (0.94%) show the lowest rate. The frequency of PMs among Middle Eastern populations, Asians, and Americans is in the range of 2–3%. *CYP2D6* gene duplications are relatively infrequent among Northern Europeans, but in East Africa the frequency of alleles with duplication of *CYP2D6* is as high as 29% [53].

CYP2D6-Related Therapeutic Response to a Multifactorial Treatment in Dementia

Few prospective clinical trials have been performed to elucidate the influence of *CYP2D6* variants on the therapeutic outcome in AD in response to cholinesterase inhibitors or other antidementia drugs. We have performed the first prospective study in AD patients who received a combination therapy with (a) an endogenous nucleotide and choline donor, CDP-choline (500 mg/day), (b) a nootropic substance, piracetam (1600 mg/day), (c) a vasoactive compound, 1,6 dimethyl 8β-(5-bromonicotinoyl-oxymethyl)-10α-methoxyergoline (nicergoline) (5 mg/day), and (d) a cholinesterase inhibitor, donepezil (5 mg/day), for one year. With this multifactorial therapeutic intervention, EMs improved their cognitive function (MMSE score) from 21.58 ± 9.02 at baseline to 23.78 ± 5.81 after 1-yr treatment. IMs also improved from 21.40 ± 6.28 to 22.50 ± 5.07 ($r = +0.96$), whereas PMs and UMs deteriorate from 20.74 ± 6.72 to 18.07 ± 5.52 ($r = -0.97$), and from 22.65 ± 6.76 to 21.28 ± 7.75 ($r = -0.92$), respectively. According to these results, PMs and UMs were the worst responders, showing a progressive cognitive decline with no therapeutic effect, and EMs and IMs were the best responders, with a clear improvement in cognition after one year of treatment. Among EMs, AD patients harboring the *1/*10 genotype responded better than patients with the *1/*1 genotype. The best responders among IMs were the *1/*3, *1/*6 and *1/*5 genotypes, whereas the *1/*4, *10/*10, and *4/*10 genotypes were poor responders. Among PMs and UMs, the poorest responders were carriers of the *4/*4 and *1×N/*1 genotypes, respectively [4,9,22,25,36–39]. In a recent study, Pilotto et al. [54] have confirmed the influence of *CYP2D6* variants (rs1080985) on the efficacy of donepezil in AD.

From all these data we can conclude the following: (i) The most frequent *CYP2D6* variants in the Southern European population (Iberian peninsula) are the *1/*1 (57.84%), *1/*4 (22.78%), *1×N/*1 (6.10%), *4/*4 (2.56%), and *1/*3 (2.01%) genotypes, accounting for more than 80% of the population; (ii) the frequency of EMs, IMs, PMs, and UMs is about 59.51%, 29,78%, 4.46%, and 6.23%, respectively, in the general population (GP), and 57.76, 31.05%, 5.27%, and 5.90%, respectively in AD cases; (iii) EMs are more prevalent in GP (59.51%) than in AD (57.76%); IMs are more frequent in AD (31.05%) than in GP (29.78%); the

© 2010 Blackwell Publishing Ltd

frequency of PMs is slightly higher in AD (5.27%%) than in GP (4.46%); and UMs are more frequent in GP (6.23%) than in AD (5.90%); (iv) there are differences between females and males in the distribution and frequency of *CYP2D6* genotypes which might be of relevance in therapeutic terms and risk of ADRs; (v) there is an accumulation of AD-related genes of risk in PMs and UMs; (vi) PMs and UMs tend to show higher transaminase activities than EMs and IMs; (vii) EMs and IMs are the best responders, and PMs and UMs are the worst responders to a combination therapy with cholinesterase inhibitors, neuroprotectants, and vasoactive substances; and (viii) the pharmacogenetic response in AD appears to be dependent upon the networking activity of genes involved in drug metabolism and genes involved in AD pathogenesis [4,9,22,24,25,36–39].

CYP Clustering in Alzheimer's disease

Since over half of the available drugs are metabolized via different CYP enzymes and other metabolic pathways, it is convenient to understand the networking activity of *CYP* genes and the genomic profiles of these genes in particular groups of risk. In the case of dementia, 73.71% of AD patients are *CYP2C19*-EMs, 25.12% IMs, and 1.16% PMs. The distribution and frequency of *CYP2C9* genotypes is as follows: *1/*1-EM 60.87%, *1/*2-IM 23.98%, *1/*3-IM 10.17%, *2/*2-PM 2.54%, *2/*3-PM 2.16%, and *3/*3-PM 0.25%, globally representing 60.87% *CYP2C9*-EMs, 34.16% IMs, and 4.97% PMs [23]. This is especially important because the *CYP2C9*-Ile359Leu (*CYP2C9*3* allele) and *CYP2C9*-Arg144Cys (*CYP2C9*2* allele) variants are associated with warfarin sensitivity. Clustering together *CYP2C9* and *VKORC1* variants, we can estimate that approximately 30% of the elderly population is sensitive to warfarin anticoagulants.

Concerning *CYP3A4/5* polymorphisms, 82.75% of AD cases are EMs (*CYP3A5*3/*3*), 15.88% are IMs (*CYP3A5*1/*3*), and 1.37% are UMs (*CYP3A5*1/*1*) [25].

The human *CYP3A* subfamily plays a dominant role in the metabolic elimination of more drugs than any other biotransformation enzyme. CYP3A enzyme is localized in the liver and small intestine and thus contributes to first-pass and systemic metabolism. CYP3A expression varies as much as 40-fold in liver and small intestine donor tissues. Unlike other human P450s (*CYP2D6*, *CYP2C19*) there is no evidence of a 'null' allele for *CYP3A4*. Over 50 SNPs have been identified in the *CYP3A4* gene. The most common variant, *CYP3A4*1B*, is an A-392G transition in the 5′-flanking region with an allele frequency ranging from 0% (Chinese and Japanese) to 45% (African-Americans). *CYP3A5* is polymorphically expressed in adults with readily detectable expression in about 10–20% in Caucasians, 33% in Japanese and 55% in African-Americans. The primary causal mutation for its polymorphic expression (*CYP3A5*3*) confers low CYP3A5 protein expression as a result of improper mRNA splicing and reduced translation of a functional protein. The *CYP3A5*3* allele frequency varies from approximately 50% in African-Americans to 90% in Caucasians. Functionally, microsomes from a *CYP3A5*3/*3* liver contain very low CYP3A5 protein and display on average reduced catalytic activity towards midazolam. Additional intronic or exonic mutations (*CYP3A5*5, *6*, and *7) may alter splicing and result in premature stop codons or exon deletion. As CYP3A5 is the primary extrahepatic CYP3A isoform, its polymorphic expression may be implicated in disease risk and the metabolism of endogenous steroids or xenobiotics [55].

The construction of a genetic map integrating the most prevalent *CYP2D6+CYP2C19+CYP2C9* polymorphic variants in a trigenic cluster yields 82 different haplotype-like profiles. The most frequent trigenic genotypes in the AD population are *1*1-*1*1-*1*1 (25.70%), *1*1-*1*2-*1*2 (10.66%), *1*1-*1*1-*1*1 (10.45%), *1*4-*1*1-*1*1 (8.09%), *1*4-*1*2-*1*1 (4.91%), *1*4-*1*1-*1*2 (4.65%), and *1*1-*1*3-*1*3 (4.33%). These 82 trigenic genotypes represent 36 different pharmacogenetic phenotypes. According to these trigenic clusters, only 26.51% of the patients show a pure 3EM phenotype, 15.29% are 2EM1IM, 2.04% are pure 3IM, 0% are pure 3PM, and 0% are 1UM2PM (the worst possible phenotype) [25].

Pharmacogenomics of AD-Related Genes

APOE- and *ACE*-related pharmacogenomics

The pharmacogenomics of AD is still in a very primitive stage. In over 100 clinical trials for dementia, *APOE* has been used as the only gene of reference for the pharmacogenomics of AD [4,24,25,36–38,56,57]. Several studies indicate that the presence of the *APOE-4* allele differentially affects the quality and extent of drug responsiveness in AD patients treated with cholinergic enhancers (tacrine, donepezil, galantamine, rivastigmine), neuroprotective compounds (nootropics), endogenous nucleotides (CDP-choline), immunotrophins (anapsos), neurotrophic factors (cerebrolysin), rosiglitazone, or combination therapies [4,24,25,36–39,56–58]; however, controversial results are frequently found due to methodological problems, study design, and patient recruitment in clinical trials.

In long-term open clinical trials with a multifactorial treatment, *APOE-4/4* carriers are the worst responders [4,24,25,36–39]. With a similar therapeutic protocol, *PSEN1-1/1* homozygotes are the worst responders and *PSEN1-2/2* carriers are the best responders [25]. Significant *ACE*-related therapeutic responses to multifactorial treatments have also been reported [22,39]. Among *ACE-I/D* variants, *ACE-D/D* patients were the worst responders (r = –0.58), and *ACE-I/D* carriers were the best responders (r = +0.26), with *ACE-I/I* showing an intermediate positive response (r = +0.01) [22,39]. *ACE*-related biochemical and hemodynamic phenotypes have been studied in patients with AD [4,22,23]. *ACE-I/I* patients tend to be younger than *ACE-I/D* or *ACE-D/D* patients at the time of diagnosis and also to show a more severe cognitive deterioration. Serum ApoE, total cholesterol, LDL-cholesterol, HDL-cholesterol, nitric oxide, histamine, and ACE levels are higher in *ACE-I/I* carriers than in patients with the other genotypes; in contrast, serum triglyceride and VLDL levels are notably lower in *ACE-I/I* patients compared to patients harboring the *ACE-I/D* or *ACE-D/D* genotypes, whereas Aβ levels do not show any clear difference among *ACE*-related genotypes. Cerebrovascular function tends to be worse in *ACE-D/D*, with lower brain blood flow velocities and higher pulsatility and resistance indices, than in *ACE-I/D* (intermediate cerebrovascular hemodynamics) or *ACE-I/I* (almost normal cerebrovascular function) [4,22,23,39]. The correlation between lipid levels and brain hemodynamics is very similar in this study to data observed in that of *CYP2D6*-related metabolizer profiles in which EM patients with moderate cholesterol and lipoprotein levels (as well as relatively high nitric oxide, histamine, ACE, and ApoE levels) tend to show a better cerebrovascular hemodynamic profile than AD patients with lower cholesterol and lipoprotein levels [39]. This apparently paradoxical correlation appears to indicate that major influences in cerebrovascular homeostasis and hemodynamic brain blood flow are cholesterol, lipoproteins, nitric oxide, ACE, and histamine, among many other factors, in AD, and that peripheral levels of Aβ are indifferent in this concern. On the other hand, it seems likely that low triglyceride levels may facilitate cerebrovascular function. It is also worth mentioning that *ACE-I/I* patients with the highest cholesterol levels are the worst in mental performance. Other interpretation of these data might suggest an association between poor cerebrovascular function with *ACE-D/D* and *ACE-I/D*, and an association between alterations in lipid metabolism in *ACE-I/I* [22,39].

Both *APOE* and *ACE* variants also affect behavior and the modification of behavioral changes (mood, anxiety) in dementia after nonpsychotropic pharmacological treatment [4,22,24,37,40]. At baseline, all *APOE* variants show similar anxiety and depression rates, except the *APOE-4/4* carriers who differed from the rest in significantly lower rates of anxiety and depression. Remarkable changes in anxiety were found among different *APOE* genotypes. Practically, all *APOE* variants responded with a significant diminution of anxiogenic symptoms, except patients with the *APOE-4/4* genotype who only showed a slight improvement. The best responders were *APOE-2/4* ($r = -0.87$) > *APOE-2/3* ($r = -0.77$) > *APOE-3/3* ($r = -0.69$) > *APOE-3/4* carriers ($r = -0.45$). The potential influence of *APOE* variants on anxiety and cognition in AD does not show a clear parallelism, suggesting that other more complex mechanisms are involved in the onset of anxiety in dementia. Concerning depression, all *APOE* genotypes improved their depressive symptoms with treatment except those with the *APOE-4/4* genotype, which worsen along the treatment period. The best responders were *APOE-2/4* ($r = -0.85$) > *APOE-2/3* ($r = -0.77$) > *APOE-3/3* ($r = -0.73$) > *APOE-3/4* ($r = -0.16$), and the worst responder was *APOE-4/4* ($r = +0.31$) [22,39]. Patients with each one of the 3 *ACE-I/D* indel variants are equally anxiogenic and depressive at baseline and all of them respond favorably to the multifactorial protocol by gradually reducing anxiety and depressive symptoms over the 12-month treatment period. The best responders were *ACE-I/D* ($r = -0.89$) > *ACE-D/D* ($r = -0.68$) > *ACE-I/I* ($r = -0.08$). Depressive symptoms were also similarly improved in all *ACE-I/D* variants. The best responders were *ACE-I/D* ($r = -0.88$) > *ACE-D/D* ($r = -0.55$) > *ACE-I/I* ($r = -0.13$). Comparatively, the worst responders among *ACE-I/D* variants were carriers of the *ACE-I/I* genotype which were also the poorest responders in anxiety and cognition [22,39,41].

The combination of *APOE* and *ACE* polymorphic variants in bigenic clusters yields different anxiety and depression patterns at baseline and after one year of treatment. The most anxiogenic patients at baseline are those with the *23DD*, *44ID*, and *34II* genotypes, and the least anxiogenic patients are those harboring the *23II*, *44DD*, and *23ID* genotypes. The most depressive clusters at baseline are those harboring the *23DD*, *33ID*, and *33II* genotypes, with a clear accumulation of *APOE-3/3* carriers in these groups, and the least depressive clusters are those represented by carriers of the *23II*, *44ID*, and *23ID* genotypes. All bigenic clusters show a positive anxiolytic and anti-depressive response to the multifactorial treatment, except *44DD* carriers who exhibited the worst response [22,39,41].

Influence of *APOE-CYP2D6* interaction on AD pharmacogenomics

APOE influences liver function and CYP2D6-related enzyme activity probably via regulation of hepatic lipid metabolism. It has been observed that APOE may influence liver function and drug metabolism by modifying hepatic steatosis and transaminase activity. There is a clear correlation between APOE-related TG levels and GOT, GPT, and GGT activities in AD [22,39]. Both plasma TG levels and transaminase activity are significantly lower in AD patients harboring the *APOE-4/4* genotype, probably indicating (a) that low TG levels protect against liver steatosis, and (b) that the presence of the *APOE-4* allele influences TG levels, liver steatosis, and transaminase activity. Consequently, it is very likely that APOE influences drug metabolism in the liver through different mechanisms, including interactions with enzymes such as transaminases and/or cytochrome P450-related enzymes encoded in genes of the *CYP* superfamily [22,39,41].

When *APOE* and *CYP2D6* genotypes are integrated in bigenic clusters and the *APOE+CYP2D6*-related therapeutic response to a combination therapy is analyzed in AD patients, it becomes clear that the presence of the *APOE-4/4* genotype is able to convert pure *CYP2D6*1/*1* EMs into full PMs, indicating the existence of a powerful influence of the *APOE-4* homozygous genotype on the drug metabolizing capacity of pure *CYP2D6*-EMs. In addition, a clear accumulation of *APOE-4/4* genotypes is observed among *CYP2D6* PMs and UMs [25].

Conclusions

From these studies we can conclude the following: (i) Most studies with acetylcholinesterase inhibitors indicate that the presence or absence of the *APOE-4* allele influences the therapeutic outcome in patients with AD. (ii) Multifactorial treatments combining neuroprotectants, endogenous nucleotides, nootropic agents, vasoactive substances, cholinesterase inhibitors, and NMDA antagonists associated with metabolic supplementation on an individual basis adapted to the phenotype of the patient may be useful to improve cognition and slow-down disease progression in AD. (iii) The therapeutic response in AD seems to be genotype-specific under different pharmacogenomic conditions. (iv) In monogenic-related studies, patients harboring the *APOE-4/4* genotype are the worst responders. (v) *APP*, *PSEN1* and *PSEN2* mutations influence the therapeutic response in AD. (vi) In trigenic-related studies (*APOE+PSEN1+PSEN2*) the best responders are those patients carrying the *331222–*, *341122–*, *341222–*, and *441112–* genomic clusters. (vii) The worst responders in all genomic clusters are patients with the *441122+* genotype. (viii) The interaction of several AD-related genes seems to be determinant for drug efficacy and safety.

(ix) *APOE-CYP2D6* interactions might influence the therapeutic response in AD via changes in lipid metabolism and liver function. (x) *APOE* may also interact with PSEN1, ACE, A2M, and other genes (e.g., *CETP*, *AGT*, and *NOS3*) to regulate the effect of drugs on cognition and behavioral changes in dementia. (xi) The *APOE-4/4* genotype seems to accelerate neurodegeneration anticipating the onset of the disease by 5–10 years; and in general, *APOE-4/4* carriers show a faster disease progression and a poorer therapeutic response to all available treatments than any other polymorphic variant. (xii) Pharmacogenomic studies using monogenic, bigenic, trigenic, tetragenic or polygenic clusters as a harmonization procedure to reduce genomic heterogeneity in clinical trials are very useful in order to widen the therapeutic scope of limited pharmacological resources [4,22–31,36–41].

Taking into consideration the data available, it might be inferred that at least 10–15% of the AD population may exhibit an abnormal metabolism of cholinesterase inhibitors and/or other drugs, which undergo oxidation via CYP2D6-related enzymes. Approximately 50% of this population cluster would show an ultrarapid metabolism, requiring higher doses of cholinesterase inhibitors in order to reach a therapeutic threshold, whereas the other 50% of the cluster would exhibit a poor metabolism, displaying potential adverse events at low doses. If we take into account that approximately 60–70% of therapeutic outcomes depend upon pharmacogenomic criteria (e.g., pathogenic mechanisms associated with AD-related genes), it can be postulated that pharmacogenetic and pharmacogenomic factors are responsible for 75–85% of the therapeutic response (efficacy) in AD patients treated with conventional drugs [4,24,25,36–38]. Of particular interest are the potential interactions of cholinesterase inhibitors with other drugs of current use in patients with AD, such as antidepressants, neuroleptics, antiarrhythmics, analgesics, and antiemetics, which are metabolized by the cytochrome P450 *CYP2D6* enzyme. Although most studies predict the safety of donepezil and galantamine, as the two principal cholinesterase inhibitors metabolized by CYP2D6-related enzymes few pharmacogenetic studies have been performed so far on an individual basis to personalize the treatment, and most studies reporting safety issues are the result of pooling together pharmacological and clinical information obtained with routine procedures. In certain cases, genetic polymorphism in the expression of *CYP2D6* is not expected to affect the pharmacodynamics of some cholinesterase inhibitors because major metabolic pathways are glucuronidation, O-demethylation, N-demethylation, N-oxidation, and epimerization. However, excretion rates are substantially different in EMs and PMs. For instance, in EMs, urinary metabolites resulting from O-demethylation of galantamine represent 33.2% of the dose compared with 5.2% in PMs, which show correspondingly higher urinary excretion of unchanged galantamine and its N-oxide [59]. Therefore, there are still many unanswered questions regarding the metabolism of cholinesterase inhibitors and their interaction with other drugs (potentially leading to ADRs) which require pharmacogenetic elucidation. It is also worth mentioning that dose titration (a common practice in AD patients treated with cholinesterase inhibitors, e.g., tacrine, donepezil) is an unwise strategy, since approximately 30–60% of drug failure or lack of therapeutic efficacy (and/or ADR manifestation) is not a matter of drug dosage but a problem of poor metabolizing capacity in PMs. As a general rule, it is recommended to avoid AChEIs (donepezil, galantamine) and major psychotropic drugs (neuroleptics, antidepressants) in *APOE-4/4* carriers and *CYP2D6*-PMs. In these particular cases (> 15%), alternative treatments should be administered.

To achieve a mature discipline of pharmacogenetics and pharmacogenomics in CNS disorders and dementia it would be convenient to accelerate the following processes: (a) to educate physicians and the public on the use of genetic/genomic screening in the daily clinical practice; (b) to standardize genetic testing for major categories of drugs; (c) to validate pharmacogenetic and pharmacogenomic procedures according to drug category and pathology; (d) to regulate ethical, social, and economic issues; and (e) to incorporate pharmacogenetic and pharmacogenomic procedures to both drugs in development and drugs on the market in order to optimize therapeutics [4,22–31,37–39,41].

Conflict of Interest

No potential conflict of interest relevant to this article is declared.

References

1. Cacabelos R. Psychogeriatric research. A conceptual introduction to geriatric neuroscience. *Psychogeriatrics* 2001;**1**:158–188.
2. Loveman E, Green C, Kirby J, Takeda A, Picot J, Payne E and Clegg A. The clinical and cost-effectiveness of donepezil, rivastigmine, galantamine and memantine for Alzheimer's disease. *Health Technol Assess* 2006;**10**:1–176.
3. Cacabelos R, Alvarez XA, Lombardi V, et al. Pharmacological treatment of Alzheimer disease: From phychotropic drugs and cholinesterase inhibitors to pharmacogenomics. *Drugs Today* 2000;**36**:415–499.
4. Cacabelos R. Molecular genetics of Alzheimer's disease and aging. *Meth Find Exper Clin Pharmacol* 2005;**27**(Suppl A):1–573.
5. www.ncbi.nlm.nih.gov/OMIM [Assessed 3 March 2010].
6. www.alzgene.org [Assessed 3 March 2010].
7. Selkoe DJ, Podlisny MB. Deciphering the genetic basis of Alzheimer's disease. *Annu Rev Genomics Hum Genet* 2002;**3**:67–99.
8. Suh, Y-H, Checler F. Amyloid precursor protein, presenilins, and α-synuclein: Molecular pahogenesis and pharmacological applications in Alzheimer's disease. *Phamacol Rev* 2002;**54**:469–525.
9. Cacabelos R. Donepezil in Alzheimer's disease: From conventional trials to pharmacogenetics. *Neuropsychiat Dis Treat* 2007;**3**:303–333.
10. Rogaeva E, Meng Y, Lee JH, et al. The neuronal sortilin-related receptor SORL1 is genetically associated with Alzheimer's disease. *Nat Genet* 2007;**39**:168–177.
11. Meng Y, Lee JH, Cheng R, et al. Association between SORL1 and Alzheimer disease in a genome-wide study. *Neuroreport* 2007;**18**:1761–1764.
12. Shibata N, Ohnuma T, Baba H, et al. Genetic association between SORL1 polymorphisms and Alzheimer's disease in a Japanese population. *Dement Geriatr Gogn Disord* 2008;**26**:161–164.
13. Wang Y, Rogers PM, Stayrook KR, et al. The selective Alzheimer's disease indicator-1 gene (Seladin-1/DHCR24) is a liver X receptor target gene. *Mol Pharmacol* 2008;**74**:1716–1721.
14. Peri A, Serio M. Neuroprotective effects of the Alzheimer's disease gene seladin-1. *J Mol Endocrinol* 2008;**41**:251–261.
15. Schaffer BA, Bertram L, Miller BL, et al. Association of GSK3B with Alzheimer's disease and frontotemporal dementia. *Arch Neurol* 2008;**65**:1368–1374.
16. Hsu WC, Wang HK, Lee LC, et al. Promoter polymorphisms modulating HSPA5 expression may increase susceptibility to Taiwanese Alzheimer's disease. *J Neural Transm* 2008;**115**:1537–1543.
17. Chen Y, Jia L, Wei C, et al. Association between polymorphisms in the apolipoprotein D gene and sporadic Alzheimer's disease. *Brain Res* 2008;**1233**:196–202.
18. Dreses-Werringloer U, Lambert JC, Vingtdeux V, et al. A polymorphism in CALHM1 influences Ca^{2+} homeostasis, Aβ levels, and Alzheimer's disease risk. *Cell* 2008;**133**:1149–1161.
19. Harold D, Abraham R, Hollingworth P, et al. Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. *Nat Genet* 2009;**41**:1088–1093.
20. Lambert JC, Heath S, Even G, et al. Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. *Nat Genet* 2009;**41**:1094–1099.
21. Lin MT, Simon DK, Ahn CH, et al. High aggregate burden of somatic mtDNA point mutations in aging and Alzheimer's disease brain. *Human Mol Genet* 2002;**11**:133–145.

22. Cacabelos R. Pharmacogenomics in Alzheimer's disease. *Meth Mol Biol* 2008;**448**:213–357.
23. Cacabelos R, Takeda M. Pharmacogenomics, nutrigenomics and future therapeutics in Alzheimer's disease. *Drugs Future* 2006;**31**(Suppl B):5–146.
24. Cacabelos R. Pharmacogenomic biomarkers in neuropsychiatry: The path to personalized medicine in mental disorders. In: Ritsner MS, editor. *The handbook of neuropsychiatric biomarkers, endophenotypes and genes*, Vol. 4. Netherlands: Springer, 2009; 3–63.
25. Cacabelos R. Pharmacogenomics and therapeutic strategies for dementia. *Expert Rev Mol Diag* 2009;**9**:567–611.
26. Berry N, Jobanputra V, Pal H. Molecular genetics of schizophrenia: A critical review. *J Psychiatry Neurosci* 2003;**28**:415–429.
27. Kato T. Molecular genetics of bipolar disorder and depression. *Psychiatry Clin Neurosci* 2007;**61**:3–19.
28. Cacabelos R. Pharmacogenomics for the treatment of dementia. *Ann Med* 2002;**34**:357–379.
29. Cacabelos R. The application of functional genomics to Alzheimer's disease. *Pharmacogenomics* 2003;**4**:597–621.
30. Cacabelos R. Pharmacogenomics and therapeutic prospects in Alzheimer's disease. *Exp Opin Pharmacother* 2005;**6**:1967–1987.
31. Cacabelos R. Pharmacogenomics, nutrigenomics and therapeutic optimization in Alzheimer's disease. *Aging Health* 2005;**1**:303–348.
32. Cacabelos R, Fernández-Novoa L, Lombardi V, et al. Cerebrovascular risk factors in Alzheimer's disease: Brain hemodynamics and pharmacogenomic implications. *Neurol Res* 2003;**25**:567–580.
33. Mastrangelo MA, Bowers WJ. Detailed immunohistochemical characterization of temporal and spatial progression of Alzheimer's disease-related pathologies in male triple-transgenic mice. *BMC Neurosci* 2008;**9**:81.
34. Rodríguez JJ, Jones VC, Tabuchi M, et al. Impaired adult neurogenesis in the dentate gyrus of a triple transgenic Mouse model of Alzheimer's disease. *PLoS ONE* 2008;**3**:e2935.
35. Roses AD. Pharmacogenetics and drug development: The path to safer and more effective drugs. *Nat Rev Genet* 2004;**5**:645–656.
36. Cacabelos R. Pharmacogenomics and therapeutic prospect in dementia. *Eur Arch Psychiatry Clin Neurosci* 2008;**258**(Suppl 1):28–47.
37. Cacabelos R. Pharmacogenetic basis for therapeutic optimization in Alzheimer's disease. *Mol Diag Ther* 2007;**11**:385–405.
38. Cacabelos R, Llovo R, Fraile C, et al. Pharmacogenetic aspects of therapy with cholinesterase inhibitors: The role of CYP2D6 in Alzheimer's disease pharmacogenetics. *Curr Alzheimer Res* 2007;**4**:479–500.
39. Cacabelos R. Molecular pathology and pharmacogenomics in Alzheimer's disease: Polygenic-related effects of multifactorial treatments on cognition, anxiety, and depression. *Meth Find Exper Clin Pharmacol* 2007;**29**(Suppl B):1–91.
40. Cacabelos R, Fernández-Novoa L, Pichel V, et al. Pharmacogenomic studies with a combination therapy in Alzheimer's disease. In: Takeda M, Tanaka T, Cacabelos R, editors. *Molecular neurobiology of alzheimer disease and related disorders*. Basel: Karger, 2004; 94–107.
41. Cacabelos R. Pharmacogenomics in Alzheimer's disease. In: Cohen N, editor. *Pharmacogenomics and personalized medicine*. NJ: Humana Press, 2008; 317–368.
42. Thomann PA, Roth AS, Dos Santos V, et al. Apolipoprotein E polymorphism and brain morphology in mild cognitive impairment. *Dement Geriatr Cogn Disord* 2008;**26**:300–305.
43. Sando SB, Melquist S, Cannon A, et al. APOEε4 lowers age at onset and is a high risk factor for Alzheimer's disease; A case control study from central Norway. *BMC Neurology* 2008. doi:10.1186/1471-2377-8-9.
44. Giacobini E. Cholinesterases in human brain: The effect of cholinesterase inhibitors on Alzheimer's disease and related disorders. In: Giacobini E, Pepeu G, editors. *The brain cholinergic system in health and disease*. Oxon: Informa Healthcare, 2006; 235–264.
45. Reisberg B, Doody R, Stoffler A, et al. Memantine in moderate-to-severe Alzheimer's disease. *N Engl J Med* 2003;**348**:1333–1341.
46. Schenk DB, Seubert P, Grundman M, et al. Aβ immunotherapy: Lessons learned for potential treatment of Alzheimer's disease. *Neurodegener Dis* 2005;**2**:255–260.
47. Verrils NM. Clinical proteomics: Present and future prospects. *Clin Biochem Rev* 2006;**27**:99–116.
48. Cacabelos R, editor. *World guide for drug use and pharmacogenomics*. Coruña, Spain: EuroEspes Publishing, 2010, (In Press).
49. www.cypalleles.ki.se [Assessed 3 March 2010].
50. www.pharmgkb.org [Assessed 3 March 2010].
51. Ozawa S, Soyama A, Saeki M, et al. Ethnic differences in genetic polymorphisms of CYP2D6, CYP2C19, CYP3As and MDR1/ABCB1. *Drug Metab Pharmacokin* 2004;**19**:83–95.
52. Sachse C, Brockmoller J, Bauer S, et al. Cytochrome P450 2D6 variants in a Caucasian population: Allele frequencies and phenotypic consequences. *Am J Hum Genet* 1997;**60**:284–295.
53. Weinshilboum RM, Wang L. Pharmacogenetics and pharmacogenomics: Development, science, and translation. *Annu Rev Genomics Hum Genet* 2006;**7**:223–245.
54. Pilotto A, Franceschi M, D'Onofrio G, et al. Effect of a CYP2D6 polymorphism on the efficacy of donepezil in patients with Alzheimer disease. *Neurology* 2009;**73**:761–767.
55. Lamba JK, Lin YS, Schuetz EG, et al. Genetic contribution to variable human CYP3A-mediated metabolism. *Adv Drug Deliv Rev* 2002;**54**:1271–1294.
56. Roses AD. The medical and economic roles of pipeline pharmacogenetics: Alzheimer's disease as a model of efficacy and HLA-B(*)5701 as a model of safety. *Neuropsychopharmacology* 2009;**34**:6–17.
57. Roses AD, Saunders AM, Huang Y, Strum J, Weisgraber KH, Mahley RW. Complex disease-associated pharmacogenetics: Drug efficacy, drug safety, and confirmation of a pathogenic hypothesis (Alzheimer's disease). *Pharmacogenomics J* 2007;**7**:10–28.
58. Risner ME, Saunders AM, Altman JF, et al. Efficacy of rosiglitazone in a genetically defined population with mild-to-moderate Alzheimer's disease. *Pharmacogenomics J* 2006;**6**:246–254.
59. Mannens GS, Snel CA, Hendrickx J, et al. The metabolism and excretion of galantamine in rats, dogs, and humans. *Drug Metab Dispos* 2002;**30**:553–563.

Comprehensive genomic characterization defines human glioblastoma genes and core pathways

The Cancer Genome Atlas Research Network*

Human cancer cells typically harbour multiple chromosomal aberrations, nucleotide substitutions and epigenetic modifications that drive malignant transformation. The Cancer Genome Atlas (TCGA) pilot project aims to assess the value of large-scale multi-dimensional analysis of these molecular characteristics in human cancer and to provide the data rapidly to the research community. Here we report the interim integrative analysis of DNA copy number, gene expression and DNA methylation aberrations in 206 glioblastomas—the most common type of primary adult brain cancer—and nucleotide sequence aberrations in 91 of the 206 glioblastomas. This analysis provides new insights into the roles of *ERBB2*, *NF1* and *TP53*, uncovers frequent mutations of the phosphatidylinositol-3-OH kinase regulatory subunit gene *PIK3R1*, and provides a network view of the pathways altered in the development of glioblastoma. Furthermore, integration of mutation, DNA methylation and clinical treatment data reveals a link between *MGMT* promoter methylation and a hypermutator phenotype consequent to mismatch repair deficiency in treated glioblastomas, an observation with potential clinical implications. Together, these findings establish the feasibility and power of TCGA, demonstrating that it can rapidly expand knowledge of the molecular basis of cancer.

Cancer is a disease of genome alterations: DNA sequence changes, copy number aberrations, chromosomal rearrangements and modification in DNA methylation together drive the development and progression of human malignancies. With the complete sequencing of the human genome and continuing improvement of high-throughput genomic technologies, it is now feasible to contemplate comprehensive surveys of human cancer genomes. The Cancer Genome Atlas aims to catalogue and discover major cancer-causing genome alterations in large cohorts of human tumours through integrated multi-dimensional analyses.

The first cancer studied by TCGA is glioblastoma (World Health Organization grade IV), the most common primary brain tumour in adults[1]. Primary glioblastoma, which comprises more than 90% of biopsied or resected cases, arises *de novo* without antecedent history of low-grade disease, whereas secondary glioblastoma progresses from previously diagnosed low-grade gliomas[1]. Patients with newly diagnosed glioblastoma have a median survival of approximately 1 year with generally poor responses to all therapeutic modalities[2]. Two decades of molecular studies have identified important genetic events in human glioblastomas, including the following: (1) dysregulation of growth factor signalling via amplification and mutational activation of receptor tyrosine kinase (RTK) genes; (2) activation of the phosphatidylinositol-3-OH kinase (PI(3)K) pathway; and (3) inactivation of the p53 and retinoblastoma tumour suppressor pathways[1]. Recent genome-wide profiling studies have also shown remarkable genomic heterogeneity among glioblastoma and the existence of molecular subclasses within glioblastoma that may, when fully defined, allow stratification of treatment[3–8]. Albeit fragmentary, such baseline knowledge of glioblastoma genetics sets the stage to explore whether novel insights can be gained from a more systematic examination of the glioblastoma genome.

Results

Data release. As a public resource, all TCGA data are deposited at the Data Coordinating Center (DCC) for public access (http://cancergenome.nih.gov/). TCGA data are classified by data type (for example, clinical, mutations, gene expression) and data level to allow structured access to this resource with appropriate patient privacy protection. An overview of the data organization is provided in the Supplementary Methods, and a detailed description is available in the TCGA Data Primer (http://tcga-data.nci.nih.gov/docs/TCGA_Data_Primer.pdf).

Biospecimen collection

Retrospective biospecimen repositories were screened for newly diagnosed glioblastoma based on surgical pathology reports and clinical records (Supplementary Fig. 1). Samples were further selected for having matched normal tissues as well as associated demographic, clinical and pathological data (Supplementary Table 1). Corresponding frozen tissues were reviewed at the Biospecimen Core Resource (BCR) to ensure a minimum of 80% tumour nuclei and a maximum of 50% necrosis (Supplementary Fig. 1). DNA and RNA extracted from qualified biospecimens were subjected to additional quality control measurements (Supplementary Methods) before distribution to TCGA centres for analyses (Supplementary Fig. 2).

After exclusion based on insufficient tumour content ($n = 234$) and suboptimal nucleic acid quality or quantity ($n = 147$), 206 of the 587 biospecimens screened (35%) were qualified for copy number, expression and DNA methylation analyses. Of these, 143 cases had matched normal peripheral blood or normal tissue DNAs and were therefore appropriate for re-sequencing. This cohort also included 21 post-treatment glioblastoma cases used for exploratory comparisons

---

*Lists of participants and their affiliations appear at the end of the paper.

(Supplementary Table 1). Although it is possible that a small number of progressive secondary glioblastomas were among the remaining 185 cases of newly diagnosed glioblastomas, this cohort represents predominantly primary glioblastoma. Indeed, when compared with published cohorts, overall survival of the newly diagnosed glioblastoma cases in TCGA is similar to that of primary glioblastomas reported in the literature (Supplementary Fig. 3, $P = 0.2$)[9–12].

Genomic and transcriptional aberrations

Genomic copy number alterations (CNAs) were measured on three microarray platforms (Supplementary Methods) and analysed with multiple analytical algorithms[13–15] (Supplementary Fig. 4 and Supplementary Tables 2–4). In addition to the well-known alterations[3,13,14], we detected significantly recurrent focal alterations not previously reported in glioblastomas, such as homozygous deletions involving *NF1* and *PARK2*, and amplifications of *AKT3* (Fig. 1a and Supplementary Tables 2–4). Search for informative but infrequent CNAs also uncovered rare focal events, such as amplifications of *FGFR2* and *IRS2*, and deletion of *PTPRD* (Supplementary Table 4). Abundance of protein-coding genes and non-coding microRNA was also measured by transcript-specific and exon-specific probes on multiple platforms (Supplementary Methods). The resulting integrated gene expression data set showed that ~76% of genes within recurrent CNAs have expression patterns that correlate with copy number (Supplementary Table 2). In addition, single-nucleotide-polymorphism (SNP)-based analyses also catalogued copy-neutral loss of heterozygosity (LOH), with the most significant region being 17p, which contains *TP53* (Supplementary Methods).

Patterns of somatic nucleotide alterations in glioblastoma

A total of 91 matched tumour–normal pairs (72 untreated and 19 treated cases) were selected from the 143 cases for detection of somatic mutations in 601 selected genes (Supplementary Table 5). The resulting sequences, totalling 97 million base pairs (1.1 ± 0.1 million bases per sample), uncovered 453 validated non-silent somatic mutations in 223 unique genes, 79 of which contained two or more events (Supplementary Table 6; see also http://tcga-data.nci.nih.gov/docs/somatic_mutations/tcga_mutations.htm). The background mutation rates differed markedly between untreated and treated glioblastomas, averaging 1.4 versus 5.8 somatic silent mutations per sample (98 events among 72 untreated cases versus 111 among 19 treated, $P < 10^{-21}$), respectively. This difference was predominantly driven by seven hypermutated samples, as determined by frequencies of both silent and non-silent mutations (Fig. 1b, c). Four of the seven hypermutated tumours were from patients previously treated with temozolomide and three were from patients treated with CCNU (lomustine) alone or in combination (Supplementary Table 1b). A hypermutator phenotype in glioblastoma has been described in three glioblastoma specimens with *MSH6* mutations[16,17], prompting us to perform a systematic analysis of the genes involved in mismatch repair (MMR). Indeed, six of the seven hypermutated samples harboured mutations in at least one of the MMR genes *MLH1*, *MSH2*, *MSH6* or *PMS2*, as compared with only one sample among the eighty-four non-hypermutated samples ($P = 7 \times 10^{-8}$), suggesting a role of decreased DNA repair competency in these highly mutated samples derived from treated patients.

By applying a statistical analysis of mutation significance[18], we identified eight genes as significantly mutated (false discovery rate $<10^{-3}$) (Fig. 2d and Supplementary Table 6). Interestingly, 27 *TP53* mutations were detected in the 72 untreated glioblastomas (37.5%) and 11 mutations in the 19 treated samples (58%). All of those mutations clustered in the DNA binding domain, a well-known hotspot for p53 mutations in human cancers (Supplementary Fig. 5 and Supplementary Table 6). Given the predominance of primary

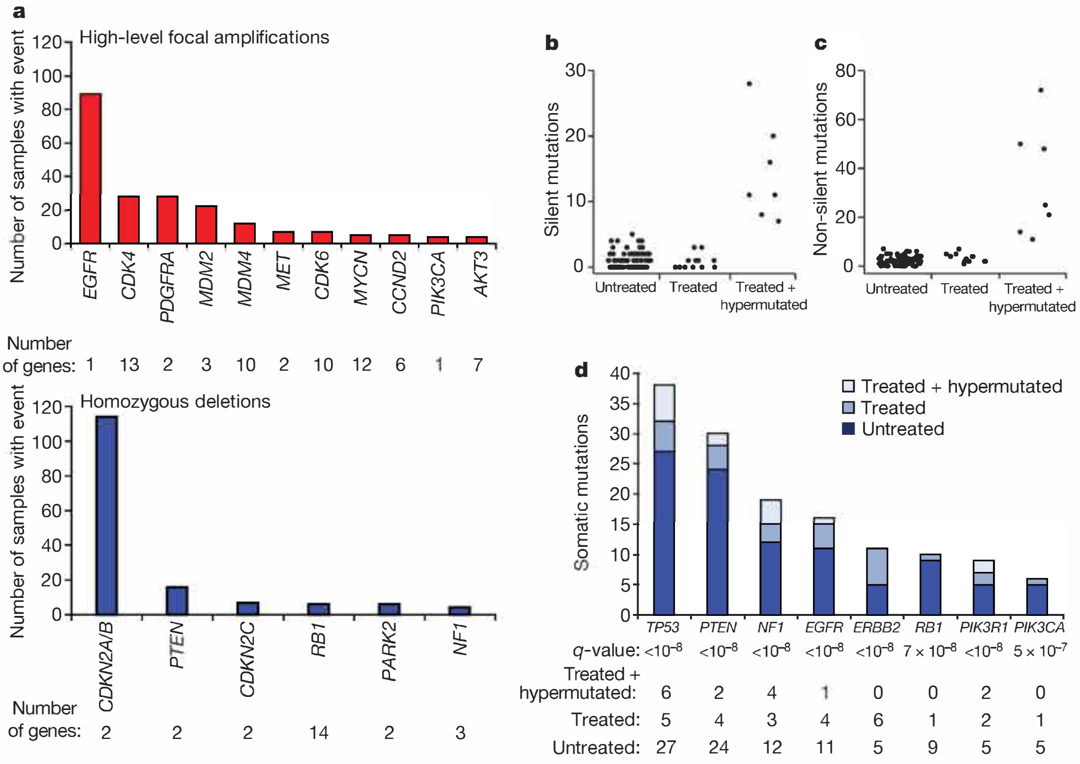


Figure 1 | Significant copy number aberrations and pattern of somatic mutations. **a**, Frequency and significance of focal high-level CNAs. Known and putative target genes are listed for each significant CNA, with 'Number of genes' denoting the total number of genes within each focal CNA boundary. **b, c**, Distribution of the number of silent (**b**) and non-silent (**c**) mutations across the 91 glioblastoma samples separated according to their treatment status, showing hypermutation in 7 out of the 19 treated samples. **d**, Significantly mutated genes in 91 glioblastomas. The eight genes attaining a false discovery rate <0.1 are displayed here. Somatic mutations occurring in untreated samples are in dark blue; those found in statistically non-hypermutated and hypermutated samples among the treated cohort are in respectively lighter shades of blue. Numbers of events in each group are noted.

glioblastoma among this newly diagnosed collection, that result unequivocally proves that p53 mutation is a common event in primary glioblastoma.

NF1 is a human glioblastoma suppressor gene. Although somatic mutations in *NF1* have been reported in a small series of human glioblastoma tumours[19], their role remains controversial[20], despite strong genetic data in mouse model systems[20–22]. Here, 19 *NF1* somatic mutations were identified in 13 samples (14% of 91), including 6 nonsense mutations, 4 splice site mutations, 5 missense changes and 4 frameshift insertions/deletions (indels) (Fig. 2a). Five of these mutations—R1391S (ref. 23), R1513* (ref. 24), e25 −1 and e29 +1 (ref. 25) and Q1966* (ref. 26)—have been reported as germline alterations in neurofibromatosis patients, and thus are probably inactivating. In addition, 30 heterozygous deletions in *NF1* were observed among the entire interim sample set of 206 cases, 6 of which also harbour point mutation (Supplementary Tables 8 and 9). Some samples also exhibited loss of expression without evidence of genomic alteration (Fig. 2b). Overall, at least 47 of these 206 patient samples (23%) harboured somatic *NF1* inactivating mutations or deletions, definitively addressing *NF1*'s relevance to sporadic human glioblastoma.

Prevalence of *EGFR* family activation. EGFR is frequently activated in primary glioblastomas. Variant III deletion of the extracellular domain ('vIII mutant')[27] has been the most commonly described event, in addition to extracellular domain point mutations and cytoplasmic domain deletions[28,29]. Here, high-resolution genomic and exon-specific transcriptomic profiling readily detected vIII and carboxy-terminal deletions with correspondingly altered transcripts (Fig. 2c). Among the 91 glioblastoma cases with somatic mutation data, 22 harboured focal amplification of wild-type *EGFR* with no point mutation, 16 had point mutations in addition to focal amplification, and 3 had *EGFR* point mutations but no amplification (Supplementary Fig. 6 and Supplementary Table 9). Collectively, *EGFR* alterations were observed in 41 of the 91 sequenced samples.

ERBB2 mutation has previously been reported in only one glioblastoma tumour[30]. In the TCGA cohort, 11 somatic *ERBB2* mutations in 7 of 91 samples were validated, including 3 in the kinase domain and 2 involving V777A, a site of recurrent missense and in-frame insertion mutations in lung, gastric and colon cancers[31]. The remaining eight mutations (including seven missense and one splice-site mutation) occurred in the extracellular domain of the protein, similar to somatic *EGFR* substitutions in glioblastoma (Fig. 2d). Unlike in breast cancers, focal amplifications of *ERBB2* were not observed in glioblastomas.

Somatic mutations of the PI(3)K complex in human glioblastoma. The PI(3)K complex consists of a catalytically active protein, p110α, encoded by *PIK3CA*, and a regulatory protein, p85α, encoded by *PIK3R1*. Frequent activating missense mutations of *PIK3CA* have been reported in multiple tumour types, including glioblastoma[32,33]. These mutations occur primarily in the adaptor binding domain (ABD) as well as the C2 helical and kinase domains[34–36]. Indeed, *PIK3CA* somatic nucleotide substitutions were detected in 6 of the 91 sequenced samples (Supplementary Table 6). Apart from the four mutations already reported in the COSMIC database (http://www.sanger.ac.uk/genetics/CGP/cosmic/), two novel in-frame deletions were detected in the adaptor binding domain of *PIK3CA* ('L10del' and 'P17del'). Those deletions may disrupt interactions between p110α and its regulatory subunit, p85α (ref. 37).

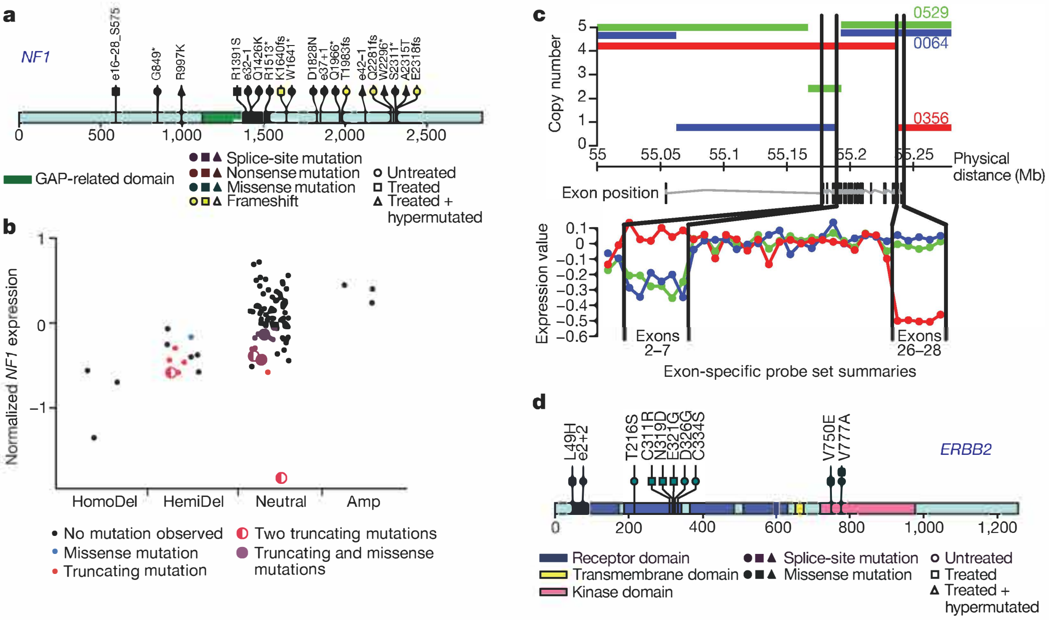


Figure 2 | Mutations in *NF1* tumour suppressor gene and *EGFR* family members. **a**, *NF1* somatic mutations in 91 glioblastoma tumours. Both missense mutations and truncating nonsense, frameshift and splice site mutations were observed. Splice positions are given in number of bases to the closest exon (e#) numbered according to the *NF1* reference transcript in the Human Gene Mutation Database; positive indicates 3′ of exon, negative indicates 5′ of exon. Asterisk indicates a stop codon. fs, frameshift. **b**, Correlation of copy number and mutation status at the *NF1* locus with level of expression (*y* axis). Mutation events predicted to result in fewer expressed copies (including deletion, nonsense, splice site and frameshift mutations) generally have lower observed expression. HomoDel, homozygous deletion; HemiDel, single-copy loss; Neutral, no change in copy number (presumed diploid); Amp, increased copy number. Copy number status of the *NF1* locus in each sample was determined as described in the Supplementary Information. **c**, DNA copy number and mRNA expression profiles for TCGA samples TCGA-08-0356 (red), TCGA-02-0064 (blue) and TCGA-02-0529 (green) at the *EGFR* locus. The upper panel shows the segmented DNA copy number (based on Affymetrix SNP6.0 data) versus genomic coordinates on chromosome 7. The lower panel shows relative exon expression levels across the known *EGFR* exons from the Affymetrix Exon array ordered by genomic position, where relative expression is the median-centred difference in exon intensity and gene intensity. The *EGFR* gene model lies between the two plots. Black lines map the genomic positions of exons 2 through to 7 and 26 through to 28. Note that structural deletions cause the relatively lower expression of exons 2–7 in the green and blue samples and exons 26–28 in the red sample. **d**, *ERBB2* somatic mutations in 91 glioblastoma tumours. Mutations cluster in the extracellular domain as seen in *EGFR* in glioblastoma. Splice site mutation position is given in number of bases to the closest exon (e#); positive indicates 3′ of exon.

Unlike *PIK3CA*, *PIK3R1* has rarely been reported as mutated in cancers. Among the five reported *PIK3R1* nucleotide substitutions in cancers[38,39], one was in a glioblastoma[39]. In our TCGA cohort, 9 *PIK3R1* somatic mutations were detected among the 91 sequenced glioblastomas. None of them was in samples with *PIK3CA* mutations. Of the nine mutations, eight lay within the intervening SH2 (or iSH2) domain and four are 3-bp in-frame deletions (Fig. 3a and Supplementary Table 6). In accord with the crystal structure of PI(3)K, which identifies the D560 and N564 amino acid residues in p85α as contact points with the N345 amino acid residue in the C2 domain of p110α (ref. 37), the mutations detected in glioblastoma cluster around those three amino acid residues (Fig. 3b), including a N345K mutation in *PIK3CA* (previously reported in colon and breast cancers[40]) and D560Y and N564K mutations in *PIK3R1*. We also identified an 18-bp deletion spanning residues D560 to S565 (DKRMNS) in *PIK3R1* (Fig. 3b) in addition to three other novel deletions (R574del, T576del and W583del) in proximity to the two key residues. We speculate that spatial constraints due to these deletions might prevent inhibitory contact of the p85α N-terminal SH2 (nSH2) domain with the helical domain of p110α, causing constitutive PI(3)K activity. Taken together, the pattern of clustering of the mutations around key residues defined by the crystal structure of PI(3)K strongly suggests that these novel *PIK3R1* point mutations and indels disrupt the important C2–iSH2 interaction, relieving the inhibitory effect of p85α on p110α.

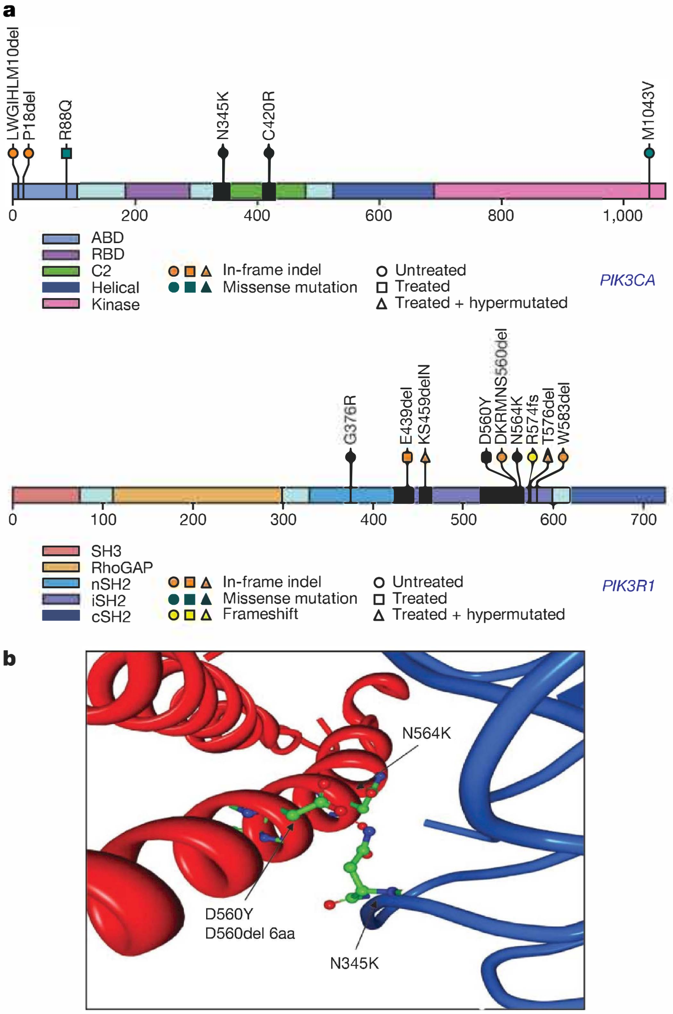

a
LWGIHLM10del
P18del
R88Q
N345K
C420R
M1043V
0
200
400
600
800
1,000
ABD
RBD
C2
Helical
Kinase
In-frame indel
Missense mutation
Untreated
Treated
Treated + hypermutated
PIK3CA
G376R
E439del
KS459delN
D560Y
DKRMNS560del
N564K
R574fs
T576del
W583del
0
100
200
300
400
500
600
700
SH3
RhoGAP
nSH2
iSH2
cSH2
In-frame indel
Missense mutation
Frameshift
Untreated
Treated
Treated + hypermutated
PIK3R1
b
N564K
D560Y
D560del 6aa
N345K

Figure 3 | *PIK3R1* and *PIK3CA* mutations in glioblastoma. **a**, The locations of mutations found in TCGA tumours are indicated above the backbone. ABD, adaptor binding domain; RBD, Ras binding domain; C2, membrane-binding domain; nSH2, N-terminal SH2 domain; iSH2, inter-SH2 domain; cSH2, C-terminal SH2 domain. **b**, Three mutations found in the interaction interface of the C2 domain of p110α with iSH2 of p85α. Two residues of p85α, D560 and N564, are within hydrogen-bonding distance of the C2 residue of p110α, N345.

MGMT methylation and MMR in treated glioblastomas

Cancer-specific DNA methylation of CpG dinucleotides located in CpG islands within the promoters of 2,305 genes was measured relative to normal brain DNA (Supplementary Table 7 and Supplementary Methods). The promoter methylation status of *MGMT*, a DNA repair enzyme that removes alkyl groups from guanine residues[41], is associated with glioblastoma sensitivity to alkylating agents[42,43]. Among the 91 sequenced cases, 19 samples were found to contain *MGMT* promoter methylation (including 13 of the 72 untreated cases and 6 of the 19 treated cases). When juxtaposed with somatic mutation data, an

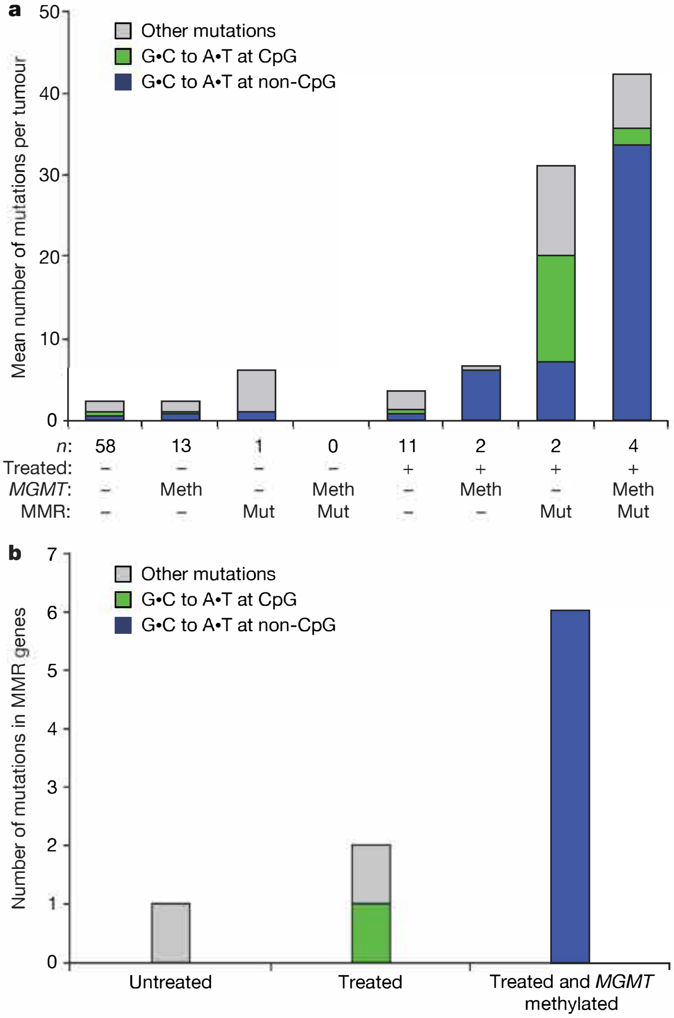


Figure 4 | Pattern of somatic mutations, *MGMT* DNA methylation and MMR gene mutations in treated glioblastomas. **a**, The mean number of validated somatic nucleotide substitutions per tumour for key sample groups is indicated on the *y* axis and denoted by the height of the bar histograms. Samples are grouped along the *x* axis according to treatment status of the patient (minus indicates untreated; plus indicates treated), DNA methylation status of *MGMT* (Meth, DNA methylated; minus, not methylated), and genetic status of MMR genes (minus, no genes mutated; Mut, one or more of the *MLH1*, *MSH2*, *MSH6* or *PMS2* genes mutated); the number below each bar indicates the number of samples in the group. Bars are colour-coded for types of nucleotide substitutions including G-to-A transitions at non-CpG sites (blue), G-to-A transitions at CpG sites (green), and other mutation types (grey). **b**, Bar histogram for mutation spectrum in the MMR genes as a function of treatment status and methylation status of *MGMT*. The colour code for substitution types is the same as in **a**.

intriguing relationship between the hypermutator phenotype and *MGMT* methylation status emerged in the treated samples. Specifically, *MGMT* methylation was associated with a profound shift in the nucleotide substitution spectrum of treated glioblastomas (Fig. 4a). Among the 13 treated samples without *MGMT* methylation, 29% (29 out of 99) of the validated somatic mutations occurred as G•C to A•T transitions in CpG dinucleotides (characteristic of spontaneous deamination of methylated cytosines), and a comparable 23% (23 out of 99) of all mutations occurred as G•C to A•T transitions in non-CpG dinucleotides. In contrast, in the six treated samples with *MGMT* methylation, 81% of all mutations (146 out of 181) turned out to be of the G•C to A•T transition type in non-CpG dinucleotides, whereas only 4% (8 out of 181) of all mutations were G•C to A•T transition mutations within CpGs. That pattern is consistent with a failure to repair alkylated guanine residues caused by treatment. In other words, *MGMT* methylation shifted the mutation spectrum of treated samples to a preponderance of G•C to A•T transition at non-CpG sites.

Notably, the mutational spectra in the MMR genes themselves reflected *MGMT* methylation status and treatment consequences. All seven mutations in MMR genes found in six *MGMT* methylated, hypermutated (treated) tumours occurred as G•C to A•T mutations at non-CpG sites (Fig. 4b and Supplementary Table 6), whereas neither MMR mutation in non-methylated, hypermutated tumours was of this characteristic. Hence, these data show that MMR deficiency and *MGMT* methylation together, in the context of treatment, exert a powerful influence on the overall frequency and pattern of somatic point mutations in glioblastoma tumours, an observation of potential clinical importance.

Integrative analyses define glioblastoma core pathways

To begin to construct an integrated view of common genetic alterations in the glioblastoma genome, we mapped the unequivocal genetic alterations—validated somatic nucleotide substitutions, homozygous deletions and focal amplifications—onto major pathways implicated in glioblastoma[1]. That analysis identified a highly interconnected network of aberrations (Supplementary Figs 7 and 8), including three major pathways: RTK signalling, and the p53 and RB tumour suppressor pathways (Fig. 5).

By copy number data alone, 66%, 70% and 59% of the 206 samples harboured somatic alterations in core components of the RB, TP53 and RTK pathways, respectively (Supplementary Table 8). In the 91 samples for which there was also sequencing data, the frequencies of somatic alterations increased to 87%, 78% and 88%, respectively (Supplementary Table 9). There was a statistical tendency towards mutual exclusivity of alterations of components within each pathway (*P*-values of 9.3×10^{-10}, 2.5×10^{-13} and 0.022, respectively, for the p53, RB and RTK pathways; Supplementary Table 10), consistent with the thesis that deregulation of one component in the pathway relieves the selective pressure for additional ones. However, we observed a greater than random chance (one-tailed, $P = 0.0018$) that a given sample harbours at least one aberrant gene from each of the three pathways (Supplementary Table 10). In fact, 74% harboured aberrations in all three pathways, a pattern suggesting that deregulation of the three pathways is a core requirement for glioblastoma pathogenesis.

As well as frequent deletions and mutations of the *PTEN* lipid phosphatase tumour suppressor gene, 86% of the glioblastoma samples harboured at least one genetic event in the core RTK/PI3K pathway (Fig. 5a). In addition to *EGFR* and *ERBB2*, *PDGFRA* (13%) and *MET* (4%) showed frequent aberrations (Supplementary Table 9). A total of 10 of the 91 sequenced samples have amplifications or point mutations in at least 2 of the 4 RTKs catalogued (*EGFR*, *ERBB2*, *PDGFRA* and *MET*; Supplementary Table 9), suggesting that genomic activation can be a mechanism for co-activated RTKs[44].

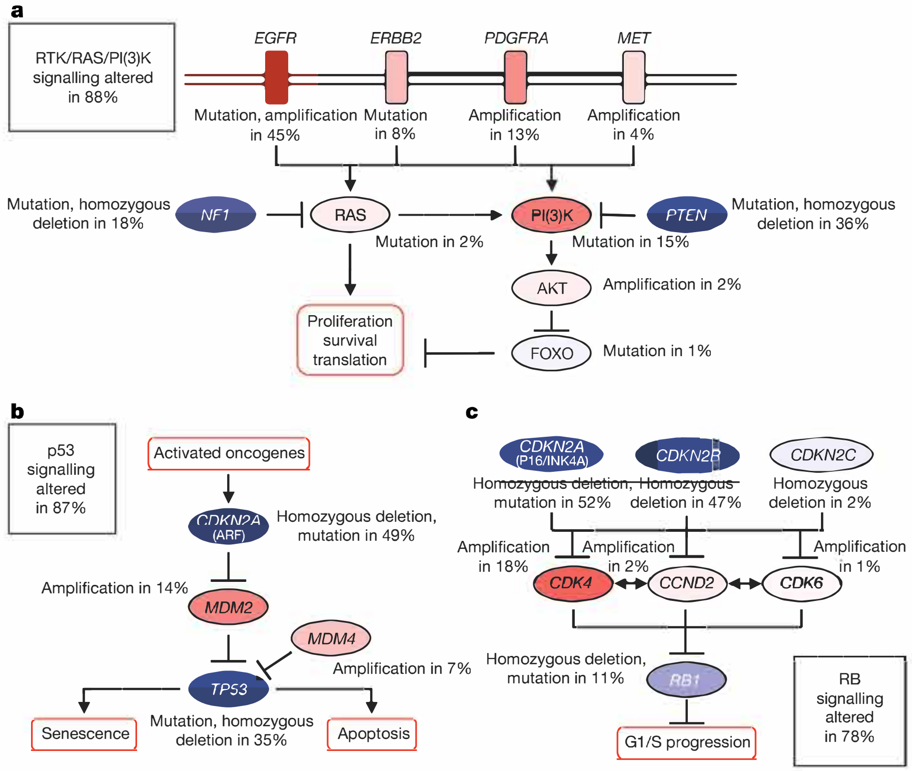

a
RTK/RAS/PI(3)K signalling altered in 88%
EGFR
ERBB2
PDGFRA
MET
Mutation, amplification in 45%
Mutation in 8%
Amplification in 13%
Amplification in 4%
Mutation, homozygous deletion in 18%
NF1
RAS
PI(3)K
PTEN
Mutation, homozygous deletion in 36%
Mutation in 2%
Mutation in 15%
AKT
Amplification in 2%
Proliferation survival translation
FOXO
Mutation in 1%
b
p53 signalling altered in 87%
Activated oncogenes
CDKN2A (ARF)
Homozygous deletion, mutation in 49%
Amplification in 14%
MDM2
MDM4
Amplification in 7%
TP53
Senescence
Mutation, homozygous deletion in 35%
Apoptosis
c
CDKN2A (P16/INK4A)
CDKN2B
CDKN2C
Homozygous deletion, mutation in 52%
Homozygous deletion in 47%
Homozygous deletion in 2%
Amplification in 18%
Amplification in 2%
Amplification in 1%
CDK4
CCND2
CDK6
Homozygous deletion, mutation in 11%
RB1
G1/S progression
RB signalling altered in 78%

Figure 5 | Frequent genetic alterations in three critical signalling pathways. **a**–**c**, Primary sequence alterations and significant copy number changes for components of the RTK/RAS/PI(3)K (**a**), p53 (**b**) and RB (**c**) signalling pathways are shown. Red indicates activating genetic alterations, with frequently altered genes showing deeper shades of red. Conversely, blue indicates inactivating alterations, with darker shades corresponding to a higher percentage of alteration. For each altered component of a particular pathway, the nature of the alteration and the percentage of tumours affected are indicated. Boxes contain the final percentages of glioblastomas with alterations in at least one known component gene of the designated pathway.

Inactivation of the p53 pathway occurred in the form of *ARF* deletions (55%), amplifications of *MDM2* (11%) and *MDM4* (4%), in addition to mutations of p53 itself (Fig. 5b and Supplementary Table 8). Among 91 sequenced samples (Supplementary Table 9), genetic lesions in *TP53* were mutually exclusive of those in *MDM2* or *MDM4* (odds ratios of 0.00 for both; $P = 0.02$ and 0.068, respectively; Supplementary Table 10), but not of those in *ARF*. In fact, 10 of the 32 tumours with *TP53* mutations also had deleted *ARF*, suggesting that homozygous deletion of the *CDKN2A* locus (which encodes both $\mathrm{p16}^{\mathrm{INK4A}}$ and ARF) was at least in part driven by $\mathrm{p16}^{\mathrm{INK4A}}$.

Among the 77% samples harbouring RB pathway aberrations (Fig. 5c), the most common event was deletion of the *CDKN2A*/*CDKN2B* locus on chromosome 9p21 (55% and 53%), followed by amplification of the *CDK4* locus (14%) (Fig. 5c and Supplementary Tables 8 and 9). Although CNAs in the CDK/RB pathway members can co-occur in the same tumour[14], all nine samples with *RB1* nucleotide substitutions (Supplementary Table 9) lacked *CDKN2A*/*CDKN2B* deletion or other CNAs in the pathway, suggesting that inactivation of *RB1* by nucleotide substitution, in contrast to copy number loss, obviates the genetic pressure for activation of upstream cyclin/cyclin-dependent kinases.

Discussion

In establishing this pilot programme, TCGA has developed important principles in biospecimen banking and collection, and established the infrastructure that will serve similar efforts in the future. Although it ensured high-quality data, the stringent biospecimen selection criteria may have introduced a degree of bias because small samples and samples with high levels of necrosis were excluded. Nonetheless, the clinical parameters of this cohort are similar to other published cohorts (Supplementary Fig. 3 and Supplementary Table 1).

The integrated analyses of multi-dimensional genomic data from complementary technology platforms have proved informative. In addition to pinpointing deregulation of RB, p53 and RTK/RAS/PI(3)K pathways as obligatory events in most, and perhaps all, glioblastoma tumours, the patterns of mutations may also inform future therapeutic decisions. It would be reasonable to speculate that patients with deletions or inactivating mutations in *CDKN2A* or *CDKN2C* or patients with amplifications of *CDK4*/*CDK6* would be candidates for treatment with CDK inhibitors, a strategy not likely to be effective in patients with *RB1* mutation. Similarly, patients with *PTEN* deletions or activating mutations in *PIK3CA* or *PIK3R1* might be expected to benefit from a PI(3)K or PDK1 inhibitor, whereas tumours in which the PI(3)K pathway is altered by *AKT3* amplification might prove refractory to those modalities. The presence of genomic co-amplification reinforces the recent report of multiple phosphorylated (activated) RTKs in individual glioblastoma specimens[44], suggesting a way to tailor anti-RTK therapeutic cocktails to specific patterns of RTK mutation. In addition, combination anti-RTK therapy might synergize with downstream inhibition of PI(3)K or cell cycle mediators. In contrast, glioblastomas with *NF1* mutations might benefit from a RAF or MEK inhibitor as part of a combination, as shown for *BRAF* mutant cancers[45].

One of the most important biomarkers for glioblastomas is the methylation status of *MGMT*, which predicts sensitivity to temozolomide[42,43], an alkylating agent that is the current standard of care for glioblastoma patients. Integrative analysis of mutation, DNA methylation and clinical (treatment) data, albeit with small sample numbers, suggests a series of inter-related events that may have an impact on clinical response and outcome. Newly diagnosed glioblastomas with *MGMT* methylation respond well to treatment with alkylating agents, in part as a consequence of unrepaired alkylated guanine residues initiating cycles of futile mismatch repair, which can lead to cell death[46–48]. Therefore, treatment of *MGMT*-deficient glioblastomas with alkylating therapy introduces a strong selective pressure to lose mismatch repair function[49]. That conclusion is consistent with our observation that the mismatch repair genes themselves are mutated with characteristic C•G to A•T transitions at non-CpG sites resulting from unrepaired alkylated guanine residues. Thus, initial methylation of *MGMT*, in conjunction with treatment, may lead to both a shift in mutation spectrum affecting mutations at mismatch repair genes and selective pressure to lose mismatch repair function. In other words, our finding raises the possibility that patients who initially respond to the frontline therapy in use today may evolve not only treatment resistance, but also an MMR-defective hypermutator phenotype. If such a hypothesis is validated, one may speculate that selective strategies designed to target mismatch-repair-deficient cells[50] would represent a rational upfront combination with alkylating agent that together may prevent or minimize emergence of such resistance. Conversely, such a treatment-mediated mutator phenotype may enhance pathway mutations that can confer resistance to targeted therapies, thereby cautioning the combination of alkylating agents with targeted agents, as this may substantially increase the probability of developing resistance to such targeted drugs.

The power of TCGA to produce unprecedented multi-dimensional data sets using statistically robust numbers of samples sets the stage for a new era in the discovery of new cancer interventions. The integrative analyses leading to the formulation of an unanticipated hypothesis on a potential mechanism of resistance highlights precisely the value and power of such project design, demonstrating how unbiased and systematic cancer genome analyses of large sample cohorts can lead to important discoveries.

METHODS SUMMARY

Biospecimens were screened from retrospective banks of tissue source sites under appropriate Institutional Review Board approvals for newly diagnosed glioblastoma with minimal 80% tumour cell percentage. RNA and DNA extracted from qualified specimens were distributed to TCGA centres for analysis. Whole-genome-amplified genomic DNA samples from tumours and normal samples were sequenced by the Sanger method. Mutations were called, verified using a second genotyping platform, and systematically analysed to identify significantly mutated genes after correcting for the background mutation rate for nucleotide type and the sequence coverage of each gene. DNA copy number analyses were performed using the Agilent 244K, Affymetrix SNP6.0 and Illumina 550K DNA copy number platforms. Sample-specific and recurrent copy number changes were identified using various algorithms (GISTIC, GTS, RAE). Messenger RNA and microRNA (miRNA) expression profiles were generated using Affymetrix U133A, AffymetrixExon 1.0 ST, custom Agilent 244K, and Agilent miRNA array platforms. mRNA expression profiles were integrated into a single estimate of relative gene expression for each gene in each sample. Methylation at CpG dinucleotides was measured using the Illumina GoldenGate assay. All data for DNA sequence alterations, copy number, mRNA expression, miRNA expression and CpG methylation were deposited in standard common formats in the TCGA DCC at http://cancergenome.nih.gov/dataportal/. All archives submitted to DCC were validated to ensure a common document structure and to ensure proper use of identifying information.

Received 28 July; accepted 1 September 2008.
Published online 4 September 2008, corrected online 17 September 2008.

1. Furnari, F. B. *et al.* Malignant astrocytic glioma: genetics, biology, and paths to treatment. *Genes Dev.* **21,** 2683–2710 (2007).
2. Mischel, P. S. & Cloughesy, T. F. Targeted molecular therapy of GBM. *Brain Pathol.* **13,** 52–61 (2003).
3. Mischel, P. S., Nelson, S. F. & Cloughesy, T. F. Molecular analysis of glioblastoma: pathway profiling and its implications for patient therapy. *Cancer Biol. Ther.* **2,** 242–247 (2003).
4. Phillips, H. S. *et al.* Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. *Cancer Cell* **9,** 157–173 (2006).
5. Maher, E. A. *et al.* Marked genomic differences characterize primary and secondary glioblastoma subtypes and identify two distinct molecular and clinical secondary glioblastoma entities. *Cancer Res.* **66,** 11502–11513 (2006).
6. Lee, J. *et al.* Tumor stem cells derived from glioblastomas cultured in bFGF and EGF more closely mirror the phenotype and genotype of primary tumors than do serum-cultured cell lines. *Cancer Cell* **9,** 391–403 (2006).
7. Diehn, M. *et al.* Identification of noninvasive imaging surrogates for brain tumor gene-expression modules. *Proc. Natl Acad. Sci. USA* **105,** 5213–5218 (2008).

8. Liang, Y. *et al.* Gene expression profiling reveals molecularly and clinically distinct subtypes of glioblastoma multiforme. *Proc. Natl Acad. Sci. USA* **102,** 5814–5819 (2005).
9. Freije, W. A. *et al.* Gene expression profiling of gliomas strongly predicts survival. *Cancer Res.* **64,** 6503–6510 (2004).
10. Murat, A. *et al.* Stem cell-related "self-renewal" signature and high epidermal growth factor receptor expression associated with resistance to concomitant chemoradiotherapy in glioblastoma. *J. Clin. Oncol.* **26,** 3015–3024 (2008).
11. Nutt, C. L. *et al.* Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. *Cancer Res.* **63,** 1602–1607 (2003).
12. Sun, L. *et al.* Neuronal and glioma-derived stem cell factor induces angiogenesis within the brain. *Cancer Cell* **9,** 287–300 (2006).
13. Beroukhim, R. *et al.* Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proc. Natl Acad. Sci. USA* **104,** 20007–20012 (2007).
14. Wiedemeyer, R. *et al.* Feedback circuit among INK4 tumor suppressors constrains human glioblastoma development. *Cancer Cell* **13,** 355–364 (2008).
15. Taylor, B. S. *et al.* Functional copy-number alterations in cancer. *PLoS ONE* (in the press).
16. Hunter, C. *et al.* A hypermutation phenotype and somatic MSH6 mutations in recurrent human malignant gliomas after alkylator chemotherapy. *Cancer Res.* **66,** 3987–3991 (2006).
17. Cahill, D. P. *et al.* Loss of the mismatch repair protein MSH6 in human glioblastomas is associated with tumor progression during temozolomide treatment. *Clin. Cancer Res.* **13,** 2038–2045 (2007).
18. Getz, G. *et al.* Comment on "The consensus coding sequences of human breast and colorectal cancers". *Science* **317,** 1500 (2007).
19. Thiel, G. *et al.* Somatic mutations in the neurofibromatosis 1 gene in gliomas and primitive neuroectodermal tumours. *Anticancer Res.* **15,** 2495–2499 (1995).
20. Zhu, Y. *et al.* Early inactivation of p53 tumor suppressor gene cooperating with NF1 loss induces malignant astrocytoma. *Cancer Cell* **8,** 119–130 (2005).
21. Reilly, K. M., Loisel, D. A., Bronson, R. T., McLaughlin, M. E. & Jacks, T. Nf1;Trp53 mutant mice develop glioblastoma with evidence of strain-specific effects. *Nature Genet.* **26,** 109–113 (2000).
22. Kwon, C. H. *et al.* Pten haploinsufficiency accelerates formation of high-grade astrocytomas. *Cancer Res.* **68,** 3286–3294 (2008).
23. Upadhyaya, M. *et al.* Mutational and functional analysis of the neurofibromatosis type 1 (NF1) gene. *Hum. Genet.* **99,** 88–92 (1996).
24. Side, L. *et al.* Homozygous inactivation of the *NF1* gene in bone marrow cells from children with neurofibromatosis type 1 and malignant myeloid disorders. *N. Engl. J. Med.* **336,** 1713–1720 (1997).
25. Fahsold, R. *et al.* Minor lesion mutational spectrum of the entire *NF1* gene does not explain its high mutability but points to a functional domain upstream of the GAP-related domain. *Am. J. Hum. Genet.* **66,** 790–818 (2000).
26. Messiaen, L. M. *et al.* Exhaustive mutation analysis of the *NF1* gene allows identification of 95% of mutations and reveals a high frequency of unusual splicing defects. *Hum. Mutat.* **15,** 541–555 (2000).
27. Humphrey, P. A. *et al.* Anti-synthetic peptide antibody reacting at the fusion junction of deletion-mutant epidermal growth factor receptors in human glioblastoma. *Proc. Natl Acad. Sci. USA* **87,** 4207–4211 (1990).
28. Lee, J. C. *et al.* Epidermal growth factor receptor activation in glioblastoma through novel missense mutations in the extracellular domain. *PLoS Med.* **3,** e485 (2006).
29. Ekstrand, A. J., Sugawa, N., James, C. D. & Collins, V. P. Amplified and rearranged epidermal growth factor receptor genes in human glioblastomas reveal deletions of sequences encoding portions of the N- and/or C-terminal tails. *Proc. Natl Acad. Sci. USA* **89,** 4309–4313 (1992).
30. Stephens, P. *et al.* Lung cancer: intragenic ERBB2 kinase mutations in tumours. *Nature* **431,** 525–526 (2004).
31. Bamford, S. *et al.* The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *Br. J. Cancer* **91,** 355–358 (2004).
32. Samuels, Y. *et al.* High frequency of mutations of the *PIK3CA* gene in human cancers. *Science* **304,** 554 (2004).
33. Gallia, G. L. *et al.* *PIK3CA* gene mutations in pediatric and adult glioblastoma multiforme. *Mol. Cancer Res.* **4,** 709–714 (2006).
34. Bader, A. G., Kang, S., Zhao, L. & Vogt, P. K. Oncogenic PI3K deregulates transcription and translation. *Nature Rev. Cancer* **5,** 921–929 (2005).
35. Bader, A. G., Kang, S. & Vogt, P. K. Cancer-specific mutations in *PIK3CA* are oncogenic *in vivo*. *Proc. Natl Acad. Sci. USA* **103,** 1475–1479 (2006).
36. Liu, Z. & Roberts, T. M. Human tumor mutants in the p110α subunit of PI3K. *Cell Cycle* **5,** 675–677 (2006).
37. Huang, C. H. *et al.* The structure of a human p110α/p85α complex elucidates the effects of oncogenic PI3Kα mutations. *Science* **318,** 1744–1748 (2007).
38. Philp, A. J. *et al.* The phosphatidylinositol 3′-kinase p85α gene is an oncogene in human ovarian and colon tumors. *Cancer Res.* **61,** 7426–7429 (2001).
39. Mizoguchi, M., Nutt, C. L., Mohapatra, G. & Louis, D. N. Genetic alterations of phosphoinositide 3-kinase subunit genes in human glioblastomas. *Brain Pathol.* **14,** 372–377 (2004).
40. Zhang, H. *et al.* Comprehensive analysis of oncogenic effects of *PIK3CA* mutations in human mammary epithelial cells. *Breast Cancer Res. Treat.* doi:10.1007/s10549-007-9847-6 (2007).
41. Pegg, A. E., Dolan, M. E. & Moschel, R. C. Structure, function, and inhibition of O6-alkylguanine-DNA alkyltransferase. *Prog. Nucleic Acid Res. Mol. Biol.* **51,** 167–223 (1995).
42. Esteller, M. *et al.* Inactivation of the DNA-repair gene MGMT and the clinical response of gliomas to alkylating agents. *N. Engl. J. Med.* **343,** 1350–1354 (2000).
43. Hegi, M. E. *et al.* *MGMT* gene silencing and benefit from temozolomide in glioblastoma. *N. Engl. J. Med.* **352,** 997–1003 (2005).
44. Stommel, J. M. *et al.* Coactivation of receptor tyrosine kinases affects the response of tumor cells to targeted therapies. *Science* **318,** 287–290 (2007).
45. Solit, D. B. *et al.* BRAF mutation predicts sensitivity to MEK inhibition. *Nature* **439,** 358–362 (2006).
46. Drablos, F. *et al.* Alkylation damage in DNA and RNA-repair mechanisms and medical significance. *DNA Repair* **3,** 1389–1407 (2004).
47. Hirose, Y., Kreklau, E. L., Erickson, L. C., Berger, M. S. & Pieper, R. O. Delayed repletion of O6-methylguanine-DNA methyltransferase resulting in failure to protect the human glioblastoma cell line SF767 from temozolomide-induced cytotoxicity. *J. Neurosurg.* **98,** 591–598 (2003).
48. Kaina, B., Christmann, M., Naumann, S. & Roos, W. P. MGMT: key node in the battle against genotoxicity, carcinogenicity and apoptosis induced by alkylating agents. *DNA Repair* **6,** 1079–1099 (2007).
49. Casorelli, I., Russo, M. T. & Bignami, M. Role of mismatch repair and MGMT in response to anticancer therapies. *Anticancer Agents Med. Chem.* **8,** 368–380 (2008).
50. Yang, J. L., Qu, X. J., Yu, Y., Kohn, E. C. & Friedlander, M. L. Selective sensitivity to carboxyamidotriazole by human tumor cell lines with DNA mismatch repair deficiency. *Int. J. Cancer* **123,** 258–263 (2008).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank the members of TCGA's External Scientific Committee, the Glioblastoma Disease Working Group (http://cancergenome.nih.gov/components) and D. N. Louis for discussions; A. Mirick, J. Melone and C. Collins for administrative coordination of TCGA activities; and L. Gaffney for graphic art. This work was supported by the following grants from the United States National Institutes of Health: U54HG003067, U54HG003079, U54HG003273, U24CA126543, U24CA126544, U24CA126546, U24CA126551, U24CA126554, U24CA126561 and U24CA126563.

Author Contributions The TCGA research network contributed collectively to this study. Biospecimens were provided by the tissue source sites and processed by the Biospecimen Core Resource. Data generation and analyses were performed by the genome sequencing centres and cancer genome characterization centres. All data were released through the Data Coordinating Center. Project activities were coordinated by the NCI and NHGRI project teams. The following TCGA investigators contributed substantively to the writing of this manuscript. **Leaders:** L. Chin and M. Meyerson. **Neuropathology:** K. Aldape, D. Bigner, T. Mikkelsen and S. VandenBerg. **Databases:** A. Kahn. **Biospecimen analysis:** R. Penny, M. L. Ferguson and D. S. Gerhard. **Copy number:** G. Getz, C. Brennan, B. S. Taylor, W. Winckler, P. Park and M. Ladanyi. **Gene expression:** K. A. Hoadley, R. G. W. Verhaak, D. N. Hayes and P. Spellman. **LOH:** D. Absher and B. A. Weir. **Sequencing:** G. Getz, L. Ding, D. Wheeler, M. S. Lawrence, K. Cibulskis, E. Mardis, Jinghui Zhang and R. K. Wilson. **TP53:** L. Donehower and D. A. Wheeler. **NF1:** W. Winckler, L. Ding and Jinghui Zhang. **EGFR:** E. Purdom and W. Winckler. **ERBB2:** W. Winckler. **PIK3R1:** L. Ding, J. Wallis and E. Mardis. **DNA methylation:** P. W. Laird, J. G. Herman, L. Ding, K. E. Schuebel, D. J. Weisenberger and S. B. Baylin. **Pathway analysis:** N. Schultz, L. Donehower, D. A. Wheeler, Jun Yao, R. Wiedemeyer, J. Weinstein and C. Sander. **General:** S. B. Baylin, R. A. Gibbs, J. Gray, R. Kucherlapati, M. Ladanyi, E. S. Lander, R. M. Myers, C. M. Perou, J. Weinstein and R. K. Wilson.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to L.C. (lynda_chin@dfci.harvard.edu) or M.M. (matthew_meyerson@dfci.harvard.edu).

The Cancer Genome Atlas Research Network

Tissue source sites: Duke University Medical School Roger McLendon[1], Allan Friedman[2], Darrell Bigner[1]; **Emory University** Erwin G. Van Meir[3,4,5], Daniel J. Brat[5,6], Gena M. Mastrogianakis[3], Jeffrey J. Olson[3,4,5]; **Henry Ford Hospital** Tom Mikkelsen[7], Norman Lehman[8]; **MD Anderson Cancer Center** Ken Aldape[9], W. K. Alfred Yung[10], Oliver Bogler[11], John N. Weinstein[57]; **University of California San Francisco** Scott VandenBerg[12], Mitchel Berger[13], Michael Prados[13]

Genome sequencing centres: Baylor College of Medicine Donna Muzny[14], Margaret Morgan[14], Steve Scherer[14], Aniko Sabo[14], Lynn Nazareth[14], Lora Lewis[14], Otis Hall[14], Yiming Zhu[14], Yanru Ren[14], Omar Alvi[14], Jiqiang Yao[14], Alicia Hawes[14], Shalini Jhangiani[14], Gerald Fowler[14], Anthony San Lucas[14], Christie Kovar[14], Andrew Cree[14], Huyen Dinh[14], Jireh Santibanez[14], Vandita Joshi[14], Manuel L. Gonzalez-Garay[14], Christopher A. Miller[14,15], Aleksandar Milosavljevic[14,15,16], Larry Donehower[17], David A. Wheeler[14], Richard A. Gibbs[14]; **Broad Institute of MIT and Harvard** Kristian

Cibulskis[18], Carrie Sougnez[18], Tim Fennell[18], Scott Mahan[18], Jane Wilkinson[18], Liuda Ziaugra[18], Robert Onofrio[18], Toby Bloom[18], Rob Nicol[18], Kristin Ardlie[18], Jennifer Baldwin[18], Stacey Gabriel[18], Eric S. Lander[18,19,20]; **Washington University in St Louis** Li Ding[21], Robert S. Fulton[21], Michael D. McLellan[21], John Wallis[21], David E. Larson[21], Xiaoqi Shi[21], Rachel Abbott[21], Lucinda Fulton[21], Ken Chen[21], Daniel C. Koboldt[21], Michael C. Wendl[21], Rick Meyer[21], Yuzhu Tang[21], Ling Lin[21], John R. Osborne[21], Brian H. Dunford-Shore[21], Tracie L. Miner[21], Kim Delehaunty[21], Chris Markovic[21], Gary Swift[21], William Courtney[21], Craig Pohl[21], Scott Abbott[21], Amy Hawkins[21], Shin Leong[21], Carrie Haipek[21], Heather Schmidt[21], Maddy Wiechert[21], Tammi Vickery[21], Sacha Scott[21], David J. Dooling[21], Asif Chinwalla[21], George M. Weinstock[21], Elaine R. Mardis[21], Richard K. Wilson[21]

Cancer genome characterization centres: Broad Institute/Dana-Farber Cancer Institute Gad Getz[18], Wendy Winckler[18,22,23], Roel G. W. Verhaak[18,22,23], Michael S. Lawrence[18], Michael O'Kelly[18], Jim Robinson[18], Gabriele Alexe[18], Rameen Beroukhim[18,22,23], Scott Carter[18], Derek Chiang[18,22], Josh Gould[18], Supriya Gupta[18], Josh Korn[18], Craig Mermel[18,22], Jill Mesirov[18], Stefano Monti[18], Huy Nguyen[18], Melissa Parkin[18], Michael Reich[18], Nicolas Stransky[18], Barbara A. Weir[18,22,23], Levi Garraway[18,22,23], Todd Golub[18,22,23], Matthew Meyerson[18,22,23]; **Harvard Medical School/Dana-Farber Cancer Institute** Lynda Chin[22,24,25], Alexei Protopopov[24], Jianhua Zhang[24], Ilana Perna[24], Sandy Aronson[26], Narayanan Sathiamoorthy[26], Georgia Ren[24], Jun Yao[24], W. Ruprecht Wiedemeyer[22], Hyunsoo Kim[26], Sek Won Kong[27,28], Yonghong Xiao[24], Isaac S. Kohane[26,27,29], Jon Seidman[30], Peter J. Park[26,27,29], Raju Kucherlapati[26]; **Johns Hopkins/University of Southern California** Peter W. Laird[31], Leslie Cope[32], James G. Herman[33], Daniel J. Weisenberger[31], Fei Pan[31], David Van Den Berg[31], Leander Van Neste[34], Joo Mi Yi[33], Kornel E. Schuebel[33], Stephen B. Baylin[33]; **HudsonAlpha Institute/Stanford University** Devin M. Absher[35], Jun Z. Li[36], Audrey Southwick[37], Shannon Brady[37], Amita Aggarwal[37], Tisha Chung[37], Gavin Sherlock[37], James D. Brooks[38], Richard M. Myers[35]; **Lawrence Berkeley National Laboratory** Paul T. Spellman[39], Elizabeth Purdom[40], Lakshmi R. Jakkula[39], Anna V. Lapuk[39], Henry Marr[39], Shannon Dorton[39], Yoon Gi Choi[41], Ju Han[39], Amrita Ray[39], Victoria Wang[40], Steffen Durinck[39], Mark Robinson[42], Nicholas J. Wang[39], Karen Vranizan[41], Vivian Peng[41], Eric Van Name[41], Gerald V. Fontenay[39], John Ngai[41], John G. Conboy[39], Bahram Parvin[39], Heidi S. Feiler[39], Terence P. Speed[40,42], Joe W. Gray[39]; **Memorial Sloan-Kettering Cancer Center** Cameron Brennan[43], Nicholas D. Socci[44], Adam Olshen[45], Barry S. Taylor[44,46], Alex Lash[44], Nikolaus Schultz[44], Boris Reva[44], Yevgeniy Antipin[44], Alexey Stukalov[44], Benjamin Gross[44], Ethan Cerami[44], Wei Qing Wang[44], Li-Xuan Qin[45], Venkatraman E. Seshan[45], Liliana Villafania[47], Magali Cavatore[47], Laetitia Borsu[48], Agnes Viale[47], William Gerald[48], Chris Sander[44], Marc Ladanyi[48]; **University of North Carolina, Chapel Hill** Charles M. Perou[49,50], D. Neil Hayes[51], Michael D. Topal[50,52], Katherine A. Hoadley[49], Yuan Qi[51], Sai Balu[52], Yan Shi[52], Junyuan Wu[52]

Biospecimen Core Resource: Robert Penny[53], Michael Bittner[54], Troy Shelton[53], Elizabeth Lenkiewicz[53], Scott Morris[53], Debbie Beasley[53], Sheri Sanders[53]

Data Coordinating Center: Ari Kahn[55], Robert Sfeir[55], Jessica Chen[55], David Nassau[55], Larry Feng[55], Erin Hickey[55]

Project teams: National Cancer Institute Anna Barker[58], Daniela S. Gerhard[58], Joseph Vockley[58], Carolyn Compton[58], Jim Vaught[58], Peter Fielding[58], Martin L. Ferguson[59], Carl Schaefer[56], Jinghui Zhang[56], Subhashree Madhavan[56], Kenneth H. Buetow[56]; **National Human Genome Research Institute** Francis Collins[60], Peter Good[60], Mark Guyer[60], Brad Ozenberger[60], Jane Peterson[60] & Elizabeth Thomson[60]

[1]Department of Pathology, [2]Department of Surgery, Duke University Medical Center, Durham, North Carolina 27710, USA. [3]Department of Neurosurgery, [4]Department of Hematology and Medical Oncology, [5]Winship Cancer Institute, [6]Department of Pathology and Laboratory Medicine, Emory University School of Medicine, Atlanta, Georgia 30322, USA. [7]Department of Neurological Surgery, [8]Department of Pathology, Henry Ford Hospital, Detroit, Michigan 48202, USA. [9]Department of Pathology, [10]Department of Neuro-Oncology, [11]Department of Neurosurgery, University of Texas M.D. Anderson Cancer Center, Houston, Texas 77030, USA. [12]Department of Pathology, [13]Department of Neurosurgery, University of California San Francisco, San Francisco, California 94143, USA. [14]Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas 77030, USA. [15]Graduate Program in Structural and Computational Biology and Molecular Biophysics, [16]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas 77030, USA. [17]Department of Molecular Virology and Microbiology, Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas 77030, USA. [18]The Eli and Edythe L. Broad Institute of Massachusetts Institute of Technology and Harvard University, Cambridge, Massachusetts 02142, USA. [19]Department of Biology, Institute of Massachusetts Institute of Technology, Cambridge, Massachusetts 02142, USA. [20]Department of Systems Biology, Harvard University, Boston, Massachusetts 02115, USA. [21]The Genome Center at Washington University, Department of Genetics, Washington University School of Medicine, St Louis, Missouri 63108, USA. [22]Department of Medical Oncology, [23]Center for Cancer Genome Discovery, [24]Center for Applied Cancer Science of the Belfer Institute for Innovative Cancer Science, Dana-Farber Cancer Institute, Boston, Massachusetts 02115, USA. [25]Department of Dermatology, Harvard Medical School, Boston, Massachusetts 02115, USA. [26]Harvard Medical School-Partners HealthCare Center for Genetics and Genomics, Boston, Massachusetts 02115, USA. [27]Informatics Program, [28]Department of Cardiology, Children's Hospital, Boston, Massachusetts 02115, USA. [29]Center for Biomedical Informatics, [30]Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA. [31]USC Epigenome Center, University of Southern California, Los Angeles, California 90089, USA. [32]Biometry and Clinical Trials Division, [33]Cancer Biology Division, The Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins University, Baltimore, Maryland 21231, USA. [34]Department of Molecular Biotechnology, Faculty of Bioscience and Engineering, Ghent University, Ghent B-9000, Belgium. [35]HudsonAlpha Institute for Biotechnology, Huntsville, Alabama 35806, USA. [36]Department of Human Genetics, University of Michigan, Ann Arbor, Michigan 48109, USA. [37]Department of Genetics, [38]Department of Urology, Stanford University School of Medicine, Stanford, California 94305, USA. [39]Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA. [40]Department of Statistics, [41]Functional Genomics Laboratory, University of California at Berkeley, Berkeley, California 94720, USA. [42]Walter and Eliza Hall Institute, Parkville, Victoria 3052, Australia. [43]Department of Neurosurgery, [44]Computational Biology Center, [45]Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, New York, New York 10065, USA. [46]Department of Physiology and Biophysics, Weill Cornell Graduate School of Medical Sciences, New York, New York 10065, USA. [47]Genomics Core Laboratory, Memorial Sloan-Kettering Cancer Center, New York, New York 10065, USA. [48]Department of Pathology, Human Oncology and Pathogenesis Program, Memorial Sloan-Kettering Cancer Center, New York, New York 10065, USA. [49]Department of Genetics, [50]Department of Pathology and Laboratory Medicine, [51]Department of Internal Medicine, Division of Medical Oncology, Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA. [52]Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA. [53]International Genomics Consortium, Phoenix, Arizona 85004, USA. [54]Computational Biology Division, Translational Genomics Research Institute, Phoenix, Arizona 85004, USA. [55]SRA International, Fairfax, Virginia 22033, USA. [56]Center For Biomedical Informatics and Information Technology, National Cancer Institute, Rockville, Maryland 20852, USA. [57]Department of Bioinformatics and Computational Biology, M.D. Anderson Cancer Center, Houston, Texas 77030, USA. [58]National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892, USA. [59]MLF Consulting, Arlington, Massachusetts 02474, USA. [60]National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA.

The Application of Genomics in Diabetes: Barriers to Discovery and Implementation

James S. Floyd[1] and Bruce M. Psaty[1,2,3]

Diabetes Care 2016;39:1858–1869 | DOI: 10.2337/dc16-0738

The emerging availability of genomic and electronic health data in large populations is a powerful tool for research that has drawn interest in bringing precision medicine to diabetes. In this article, we discuss the potential application of genomics to the prediction, prevention, and treatment of diabetes, and we use examples from other areas of medicine to illustrate some of the challenges involved in conducting genomics research in human populations and implementing findings in practice. At this time, a major barrier to the application of genomics in diabetes care is the lack of actionable genomic findings. Whether genomic information should be used in clinical practice requires a framework for evaluating the validity and clinical utility of this approach, an improved integration of genomic data into electronic health records, and the clinical decision support and educational resources for clinicians to use these data. Efforts to identify optimal approaches in all of these domains are in progress and may help to bring diabetes into the era of genomic medicine.

Many anticipated that the completion of the Human Genome Project over 10 years ago would mark the beginning of a new era of genomic medicine, in which new approaches to discovery research, disease prediction, and treatment would develop from an improved understanding of the genetic causes of human disease. In some areas of medicine, genomic discoveries have led to important new treatments. Genetic association studies have demonstrated that loss-of-function mutations in *PCSK9* result in low levels of LDL cholesterol and a reduced risk of coronary heart disease (1,2), This discovery led to a new class of drugs with dramatic lipid-lowering effects (3,4). In oncology, there has been a shift from using older drugs with broad cytotoxic effects to therapies that target specific mutations in driver genes (5), resulting in impressive reductions in mortality for some cancers (6).

Beyond the discovery of new drug targets, genomic information can be used to predict the occurrence of disease and to identify subgroups of patients for whom existing therapies or interventions will have the greatest efficacy or the least adverse effects. These are key elements of an approach that is now called precision medicine (7). Successes in oncology and other technological developments—the rapidly decreasing cost of whole-genome sequencing (8), improvements in informatics, and the widespread adoption of electronic health records (9–11)—have galvanized interest in applying various forms of big data, including genomics, to diseases such as diabetes (12). In this article, we discuss the application of genomics to diabetes, with a focus on some of the challenges involved in conducting genomics research in human populations and implementing findings in practice.

---

[1]Cardiovascular Health Research Unit and Departments of Epidemiology and Medicine, University of Washington, Seattle, WA
[2]Department of Health Services, University of Washington, Seattle, WA
[3]Group Health Research Institute, Seattle, WA

Corresponding author: James S. Floyd, jfloyd@uw.edu.

Received 11 April 2016 and accepted 16 August 2016.

© 2016 by the American Diabetes Association. Readers may use this article as long as the work is properly cited, the use is educational and not for profit, and the work is not altered. More information is available at http://www.diabetesjournals.org/content/license.

See accompanying articles, pp. 1854, 1870, 1874, 1879, 1889, 1896, 1902, 1909, and 1915.

GENOMICS IN THE PREDICTION, PREVENTION, AND DIAGNOSIS OF DIABETES

The incidence and prevalence of diabetes have doubled over the past two decades (13), and there are now about 30 million adults in the U.S. living with this condition, 95% of whom have type 2 diabetes (14). Genome-wide association (GWA) studies test hundreds of thousands or even millions of common (minor allele frequency [MAF] >5%) and low-frequency (MAF 1–5%) variants across both protein coding (exonic) and noncoding (intronic) regions of the genome. Large GWA studies have identified more than 50 genetic loci associated with various glycemic traits and at least 90 loci associated with type 2 diabetes (15–18). These genetic variants, which may explain as much as 10% of the variance in disease susceptibility, have advanced our understanding of the biology of diabetes, but each genetic locus confers only a small increase in risk. For example, the common variant from these GWA studies most strongly associated with type 2 diabetes, an intronic variant in *TCF7L2* (rs7903146), is associated with a 37% increased relative risk per copy of the variant allele (19). Rare variants (MAF <1%) and variants that are common only in specific ancestral populations have been associated with a greater increase in diabetes risk, but they account for less of the overall burden of diabetes (20–22).

Genetic risk scores (GRSs) that combine information from multiple genetic variants have been evaluated as a tool for the prediction of type 2 diabetes. Meigs et al. (23) found that a GRS with 18 variants was significantly associated with the risk of developing type 2 diabetes in the Framingham Heart Study (FHS) (odds ratio [OR] 1.12 per variant allele) and that persons in the highest out of three risk categories had an OR of 2.6 for developing type 2 diabetes compared with persons in the lowest risk category. However, this GRS did not improve the prediction of diabetes beyond traditional nongenetic risk factors (23), and the same was true for an updated GRS that included 65 variants (24). To put this into perspective, a prognostic marker with an OR of 3.0 that correctly identifies 80% of persons who will develop diabetes would incorrectly classify 60% of persons who will not develop diabetes (25); this degree of discrimination is not useful clinically (26).

Biologic pathways other than heritable changes in DNA sequence may also be important predictors of diabetes and account for some of the variability in diabetes susceptibility not explained by genetics or traditional environmental factors. For example, DNA methylation at CpG sites, a key epigenetic mechanism for the regulation of gene expression, has been associated with the risk of type 2 diabetes (27,28). Metabolomic profiles of amino acids and other small molecules may also play a role (29), particularly among younger adults (30). However, these new types of "omics" suffer from the traditional epidemiologic limitations of confounding and reverse causality and will require rigorous evaluation before their clinical validity and utility are understood.

Risk prediction tools are most useful when there are effective and safe prevention measures, which may include behavioral interventions or drug therapies. For high-risk adults, the Diabetes Prevention Program (DPP) lifestyle intervention reduces the risk of type 2 diabetes by more than half (31), and this intervention is now offered throughout the country at programs recognized by the Centers for Disease Control and Prevention. Although the highly effective DPP lifestyle intervention has few or no adverse effects, the identification of persons who benefit the most from the intervention could help to prioritize its deployment in resource-limited settings. Florez and colleagues (32,33) have evaluated whether several genetic variants associated with diabetes risk modified the effectiveness of the lifestyle intervention in the original DPP study, and they found little evidence of effect modification based on genetic risk. Some have argued that communicating genetic information on disease risk might help to motivate healthy behaviors, but current evidence does not support this claim (34). For example, in a small randomized controlled trial (RCT) of participants with type 2 diabetes who all underwent an intensive lifestyle intervention directed at weight loss, those who received information on their genetic risk for diabetes had the same self-reported motivation and adherence to the intervention as those who did not (35).

If genetic tests are not helpful in the prediction and prevention of diabetes, they could have a role in discriminating between type 1 and type 2 diabetes. The epidemic of obesity (36) has made it more difficult to distinguish diabetes type because many children and young adults with type 1 diabetes are also obese (37). Misclassification poses significant risks; an incorrect diagnosis of type 2 diabetes may lead to inappropriate treatment with oral glucose-lowering drugs, and an incorrect diagnosis of type 1 diabetes may lead to unnecessary insulin treatment. In a recent cross-sectional study of type 1 and type 2 diabetes, Oram et al. (38) evaluated a GRS that included high-risk HLA genotypes and 31 genetic loci for type 2 diabetes. They found that this GRS improved the discrimination between strictly defined type 1 and type 2 diabetes when added to clinical factors and autoimmune antibody tests, and it also helped to predict who would require insulin treatment within 3 years of diagnosis (38). One advantage of a diagnostic tool based on genotype is that, unlike islet cell antibodies, the result does not change over time. However, before this type of genetic testing can be recommended for routine use in the clinic, further evaluation in prospective studies will be necessary to demonstrate not only accurate discrimination between type 1 and type 2 diabetes but also improved use of appropriate glucose-lowering treatment.

Most cases of diabetes have multiple genetic and environmental causes and are classified according to the presumed pathophysiologic defect—autoimmune destruction of β-cells leading to insulin deficiency for type 1 diabetes and varying degrees of insulin resistance and deficiency for type 2 diabetes. In other words, the vast majority of diabetes is polygenic, and despite the growth in knowledge about the various genetic causes of diabetes in recent years, classification of individual cases into meaningful subtypes based on the underlying genetics has been difficult. On the other hand, genetic testing may be useful for the diagnosis of certain forms of diabetes caused by defects in a single gene, such as *HNF1A* mutations for maturity-onset diabetes of the young (MODY)

(39) and activating *KCNJ11* mutations for neonatal diabetes (40), both of which are highly responsive to sulfonylurea therapy. These monogenic forms of diabetes account for ~1–2% of diabetes cases (41,42), and they typically present at a young age (<25 years) and follow an autosomal dominant pattern of inheritance. Targeted genotyping could also play a role in the diagnosis of type 2 diabetes in specific populations. For example, a rare missense variant in *HNF1A* (p.E508K) that increased the risk of diabetes fivefold was present among 2% in a study of Latinos in the southern U.S. with type 2 diabetes (20); additional studies are needed to determine whether this functional variant shares the sulfonylurea-responsiveness of the *HNF1A* variants that cause MODY.

PHARMACOGENOMICS OF THERAPIES FOR TYPE 2 DIABETES

Although, among persons with diabetes, the rates of microvascular (retinopathy, neuropathy, kidney disease) and cardiovascular complications have decreased by about half over the past two decades, they still occur more often among persons with diabetes than among individuals without diabetes (43). Reducing these risks is the major goal of glucose-lowering therapy. For type 1 diabetes, the long-term benefits of intensive insulin therapy are well established (44,45). For type 2 diabetes, intensive glucose-lowering therapy prevents microvascular complications, and postrandomization follow-up data from several RCTs suggest there may be long-term cardiovascular benefits as well (46–48).

There are now several classes of medications approved for the treatment of diabetes (Table 1). Most oral therapies have similar average effects on hemoglobin A_{1c} (HbA_{1c}), but they differ in their contraindications and side effects (49,50). There is surprisingly little information about the comparative benefits and harms for different drugs (51), and treatment guidelines for type 2 diabetes (52) permit the use of most approved drug classes as second-line therapy after metformin, which is the recommended first-line therapy for most patients because of its good safety profile and potential cardiovascular benefits (53,54). There is also substantial interindividual variability in drug response (55), and many patients eventually fail to achieve recommended levels of glycemic control with their initial therapy (56,57). For example, in the U.K. General Practice Research Database, only half of patients who initiated therapy with metformin or a sulfonylurea achieved an HbA_{1c} level of <7% (58). The factors that account for this variation are not well understood (59). Because of the wide range of side effects from different therapies and because of the person-to-person variability in treatment response and adverse effects, pharmacogenomic testing for genetic variants that define subgroups of patients who are most likely to benefit from or least likely to be harmed by specific drugs is an attractive potential application of genomics in diabetes.

Pharmacogenomic studies have typically focused on candidate genes involved in pharmacokinetics (absorption, distribution, metabolism, and elimination) or pharmacodynamics (the biologic effect of a drug on its target) (60). The pharmacogenomics of oral diabetes therapies has been reviewed in detail elsewhere (61–63); selected findings for the most extensively studied drugs, sulfonylureas and metformin, are listed in Table 2. Sulfonylureas undergo metabolism by cytochrome P450 (CYP) enzyme 2C9, and loss-of-function variants in the *CYP2C9* gene have been associated with greater glucose-lowering effects and an increased risk of hypoglycemia (64,65). The genes encoding the sulfonylurea receptor, *KCNJ11* and *ABCC8*, have also been associated with increased sulfonylurea response in some studies (66,67) but not others (68,69), and in one study the association was in the opposite direction (70). Unlike sulfonylureas, metformin does not undergo metabolism by CYP enzymes; it is excreted intact by the kidneys (71). Genes encoding several transporters that facilitate the movement of metformin into the bloodstream, into target tissues, and into renal tubular cells have been associated with effects on serum levels of metformin, glucose-lowering effect, and drug intolerance (72,73).

Candidate gene studies of sulfonylureas and metformin have not typically accounted for false positives from multiple comparisons, and most findings from these studies have failed to replicate

Table 1—Classes of medications commonly used to treat diabetes

Drug classes	Oral	Average HbA_{1c} reduction	Other benefits	Adverse effects	Cost
Metformin	Yes	~1.0–1.5	Weight loss, possible cardiovascular benefit	Lactic acidosis (rare), gastrointestinal side effects	Low
Sulfonylureas	Yes	~1.0–1.5		Hypoglycemia, weight gain, potential cardiovascular risk	Low
Meglitinides	Yes	~1.0		Hypoglycemia, weight gain	Moderate
Thiazolidinediones	Yes	~1.0		Heart failure, myocardial infarction (rosiglitazone), bone loss and fractures	Moderate
α-Glucosidase inhibitors	Yes	~0.8		Flatulence, diarrhea	Moderate
Amylin analog	No	~0.6		Hypoglycemia, nausea	High
DPP-4 inhibitors	Yes	~0.6–0.8		Potential heart failure (saxagliptin)	High
GLP-1 receptor agonists	No	~1.0	Weight loss	Nausea, vomiting, diarrhea	High
SGLT2 inhibitors	Yes	~0.6–0.8	Possible cardiovascular benefit	Genitourinary infections, ketoacidosis (rare)	High
Insulin	No	Unlimited	Most potent treatment	Hypoglycemia, weight gain	Low-high

Adapted from refs. 49,50. DPP-4, dipeptidyl peptidase 4; GLP-1, glucagon-like peptide 1; SGLT2, sodium–glucose cotransporter 2.

Table 2—Selected pharmacogenomic findings for sulfonylureas and metformin

Drug	Locus	Phenotype	*N*	Effect**	Refs.
Sulfonylureas					
Pharmacokinetic	*CYP2C9*	HbA_{1c} response	1,073	0.5% absolute greater reduction in HbA_{1c} (homozygous for variant alleles)	64
		FBG response	475	No association	75
		Hypoglycemia	357	OR 5.2 for hypoglycemia (homozygous for variant alleles)	65
Pharmacodynamic	*KCNJ11/ABCC8**	HbA_{1c} and FBG response	1,268	3.5% relative greater reduction in HbA_{1c} and 7.7% relative greater reduction in FBG (homozygous for variant alleles)	66
		HbA_{1c} response	101	0.2% absolute greater reduction in HbA_{1c} (per variant allele)	67
		Insulin treatment	525	No association	68
		FBG response	228	No association	69
		HbA_{1c} response	97	Less reduction in HbA_{1c}	70
	TCF7L2	On treatment HbA_{1c} $<$7%	901	OR 1.9 for treatment failure (homozygous for variant allele)	164
		On treatment HbA_{1c} $<$7%	189	OR 1.6 for treatment failure (per variant allele)	165
Metformin					
Pharmacokinetic	*SLC22A1*	HbA_{1c} response	102	0.3% absolute lower reduction in HbA_{1c} (per variant allele)	166
		HbA_{1c} response	371	1.1% absolute lower reduction in HbA_{1c} (per variant allele)	77
		On treatment HbA_{1c} $<$7%	1,531	No association	76
		Drug intolerance	2,166	OR 2.4 for discontinuation (homozygous for variant alleles)	73
	SLC47A1	HbA_{1c} response	116	0.3% absolute lower reduction in HbA_{1c} (per variant allele)	167
		HbA_{1c} response	371	No association	77
		Risk of type 2 diabetes	2,994	Less reduction in diabetes risk	
	SLC47A2	HbA_{1c} response	253	0.1% absolute lower reduction in HbA_{1c} (any variant allele)	168
		HbA_{1c} response	371	No association	77
GWA studies	*ATM*	HbA_{1c} response, on treatment HbA_{1c} $<$7%	2,896	0.1% absolute greater reduction in HbA_{1c} and OR 1.4 for treatment success (per variant allele)	80
		HbA_{1c} response, on treatment HbA_{1c} $<$7%	1,366	No association with HbA_{1c} response, OR 1.2 for treatment success (per variant allele)	81

FBG, fasting blood glucose. *Loss-of-function variants rs757110 in *ABCC8* and rs5219 in *KCNJ11* are in near-complete linkage disequilibrium. **Effect listed only if reported as statistically significant in cited article.

(68,74–77). Unfortunately, this low replication rate is consistent with the rate in other candidate gene studies, which may be as low as 1–3% (78,79). The potential reasons include small sample sizes, heterogeneity of genetic architecture across different populations, and publication bias. In contrast, large GWA studies that use rigorous, prespecified statistical methods have produced more valid and reproducible results. For example, the only GWA study of treatment response to a diabetes therapy, conducted in a Scottish observational cohort with nearly 3,000 participants, identified a locus for glucose-lowering response to metformin near the *ATM* gene that met a stringent threshold for statistical significance (80), and this finding has replicated in other populations (81). Well-powered pharmacogenomic studies that use rigorous statistical methods may identify genetic variants that result in greater glucose-lowering effects or fewer adverse effects from other diabetes therapies.

GWA studies rely on information about the nonrandom association of alleles (linkage disequilibrium) to "tag" common and low-frequency variation throughout the genome in a subset of directly genotyped variants. Because these studies have identified only a small portion of the known heritability of most complex traits (82) and because theory predicts that rare variants are more likely to be damaging to gene function than common variants (83,84), there has been growing interest in studying rare variants. Advances in technology and informatics have made possible the high-throughput sequencing of whole exomes and whole genomes, which provide a complete assessment of both common and rare variation (85). A recent study that sequenced 202 genes encoding drug targets in 14,000 individuals identified on average one rare variant per 17 base pairs, and 90% of these variants were newly discovered (86). Another exome-sequencing study of 12 *CYP* genes, which are responsible for about 75% of all known oxidative drug metabolism, identified 1,006 variants in these genes and found that 73% were rare, a third were predicted to affect protein structure, and 9% of individuals had at least one newly discovered functional variant (87). Whether this abundance of rare variation explains some of the interindividual variability in drug response is unknown at this time, but the topic merits further evaluation.

CHALLENGES IN CONDUCTING GENOMIC RESEARCH

Study Power and Sample Size

One major barrier to genomic discovery in diabetes research has been the limited power to detect associations, which is a function of the frequency of the genetic variant, the magnitude of the effect to be detected, and the sample size (88). Large GWA studies with study populations of tens of thousands or more have helped to unravel the biology underlying many complex diseases, but most genetic loci identified by these studies have small effects. For example, the largest GWA study of type 2 diabetes, which included nearly 35,000 case and 115,000 control subjects, identified 65 genetic loci; all but the *TCF7L2* locus had ORs of 1.2 or lower per copy of the variant allele (15).

In pharmacogenomic studies, restriction to users of a particular drug further limits the available study population (89). Rare immune-mediated adverse drug reactions (ADRs) such as drug-induced liver injury or severe skin reactions may be caused by pathogenic variants with ORs of 100 or greater, and these associations can be detected with fewer than 50 cases (90–92). For diabetes therapies, rare adverse effects such as metformin-associated lactic acidosis or ketoacidosis related to sodium–glucose cotransporter 2 inhibitors may be amenable to pharmacogenomic discovery in similarly small studies.

It has been more difficult to identify pharmacogenomic associations with complex phenotypes such as myocardial infarction or stroke (93) or with quantitative traits such as QT interval prolongation or cholesterol lowering. For example, a large GWA study with over 30,000 participants from 10 observational cohorts evaluated QT interval prolongation from various drugs and failed to identify any pharmacogenomic loci at genome-wide levels of significance (94), and a GWA study with over 40,000 statin-using participants discovered and replicated two new loci for LDL cholesterol lowering, but each variant allele resulted in a relative change in LDL lowering effect of less than 2% (95). While these sorts of findings may reveal new information about human population biology, these effect sizes, which cannot be distinguished from intraindividual variation or measurement error in individual patients, are so small that the effort to genotype these variants can be safely omitted from clinical practice. As pharmacogenomic efforts move to whole-exome and whole-genome sequencing, even larger sample sizes may be required to identify associations with rare variants.

Individual pharmacogenomic variants often have small effects or fail to reach stringent thresholds of statistical significance, but they can be combined within a gene or within a pathway of several genes to identify clinically important effects. As an example, Dujic et al. (73) evaluated the relationship between metformin intolerance and four reduced-function variants in the gene *SLC22A1*, which transports metformin into the intestine and may mediate some of the gastrointestinal side effects from this drug. They found that the presence of any two reduced-function alleles increased the risk of metformin discontinuation by 2.4-fold (73).

To assess associations with the large number of variants identified by whole-exome and whole-genome sequencing, more innovative approaches are necessary. Annotation tools have been used to restrict analyses to variants that are likely to be functional, on the basis of expected protein structure, associations with gene expression levels (96,97), and information from the Encyclopedia of DNA Elements (ENCODE) project, which has systematically mapped nonprotein coding regulatory function—including transcription factor binding sites, chromatin structure, and histone modifications—throughout the genome (98). Various statistical methods have been developed that aggregate rare functional variants for gene-based association tests, which can improve the power to detect an association if there are multiple damaging variants within a gene (99,100). Data from whole-exome and whole-genome sequencing have not yet been systematically evaluated in pharmacogenomic studies, but for some complex traits these methods have been used to identify rare variants with large effects, both in new and previously unidentified loci (101).

The studies of statin response and drug-induced QT prolongation described in the preceding paragraphs represent one model for conducting genomic research in large populations: local analysis of deeply phenotyped cohorts followed by a meta-analysis of summary results in large research consortia (15,102). The increasing availability of electronic health data and a recognition of the large sample sizes required for genomic discovery research have led to the emergence of another model: biobank studies that genotype tens or hundreds of thousands of individuals and link these genetic data with participants' electronic health data to create large data repositories. Some of these studies include a baseline visit for physical measurements, the collection of specimens, and imaging tests (103,104), similar to the traditional cohort studies, and they all rely on electronic health databases for longitudinal information on health care encounters, laboratory tests, vital status, and medication use. Two of the largest biobank studies, the Million Veteran Program (105) and the UK Biobank (104), have recruited close to 500,000 individuals each.

Phenotyping With Electronic Health Data

Electronic health data have been immensely useful for research, but they have important limitations. Results from laboratory tests that are measured in the course of clinical care, such as HbA_{1c} or cholesterol levels, are likely to be recorded accurately and completely, but they may be related to the clinical indications for the tests and lack standardization across sites. Diagnosis codes associated with health care encounters, an important source of information about disease status, are assigned for clinical and billing rather than research purposes; geographic location (106), changes in reimbursement (107,108), and other factors can influence the assignment of these codes. For a given study design, different databases can produce different estimates of association between a drug exposure and a health outcome, sometimes in opposite directions (109). Nonetheless, diagnosis code–based algorithms have been used to identify some acute diabetes complications, such as hypoglycemia (110,111) and myocardial infarction

(112,113), with a reasonably high degree of accuracy (positive predictive value [PPV] 80–90%).

For other diabetes complications, such as heart failure, the accuracy of a diagnostic algorithm may vary substantially by the diagnosis code used and even the position of the code (primary vs. secondary) (114,115). The use of low-PPV algorithms can attenuate estimate of associations toward the null (116), sometimes dramatically. When a new diagnosis code was introduced for rhabdomyolysis, a severe ADR related to statin use, we evaluated the accuracy of this code in a large health care system by reviewing electronic medical records to validate potential cases, and we estimated the risk of rhabdomyolysis associated with the 80-mg dose of simvastatin, which was known to be high from a recent RCT (117). The PPV of the rhabdomyolysis code for the statin-related ADR was only 8%. Moreover, the relative risk for rhabdomyolysis associated with high-dose versus low-dose simvastatin was 12.2 when validated cases were evaluated, replicating the RCT estimate, but only 1.8 when the diagnosis code was used without validation (118). This marked attenuation of a genuine association is not an isolated finding; the quality (119) and severity (120) of disease phenotypes have also been shown to impact the magnitude of pharmacogenomic associations. For studies that evaluate genetic associations with disease outcomes, including ADRs, it may sometimes be necessary to conduct validation studies to assess the accuracy of claims-based data (121). Despite these limitations, the linkage of genomic data with electronic health data in large study populations holds promise as a powerful tool for research.

FRAMEWORK FOR EVALUATING A GENOMIC TEST

For the regulatory approval of new therapies directed at specific genetic defects, such as ivacaftor for the treatment of cystic fibrosis caused by a specific *CFTR* mutation (122), the U.S. Food and Drug Administration (FDA) requires the same standard used to evaluate all new drugs: substantial evidence of efficacy and safety from well-controlled studies, typically rigorously conducted RCTs. For therapies already in use that have demonstrated efficacy and safety in nongenetically determined populations, the evidence standard for using genomic information to select the best drug or dose is less certain. The FDA currently includes pharmacogenomic information in the prescribing information for over 160 drugs (123), but its oversight of laboratory-developed genetic tests is concerned primarily with analytic validity (i.e., ability to reliably measure a genetic variant or a biomarker) (124), which is inadequate to guide treatment decisions. A recent study of pharmacogenomic information in drug labels found convincing evidence of clinical validity (i.e., how accurately and consistently genetic variation predicts a phenotype) for only 36% of these drugs, and evidence of clinical utility (i.e., ability to improve outcomes for patients) for only 15% (125). The Centers for Disease Control and Prevention Office of Public Health Genomics has developed a more comprehensive framework for the evaluation of a genetic test that includes questions about analytic validity, clinical validity, and clinical utility, as well as ethical, legal, and social implications (ACCE), which are summarized in Table 3 (126,127). These standards are important to protect the health of the public.

As described earlier, most reported pharmacogenomic findings fail to replicate and therefore lack clinical validity. Similarly, rare variants identified by sequencing and initially identified as “pathogenic” are often downgraded to “uncertain significance” or “benign” after further study (128,129). For example, in the FHS and the Jackson Heart Study (JHS), seven genes responsible for MODY were sequenced and rare variants were identified that had previously been identified as causal for MODY or were predicted to be damaging to protein function by using annotation tools. These variants, present among 2% of persons in these population-based cohorts, were not associated with the risk of diabetes, and only one variant carrier out of 68 met the criteria for MODY (130).

Of the genomic findings for diabetes that have clinical validity, most have effects on disease outcomes or surrogate end points that are uncertain or too small to be clinically meaningful. Genetic risk panels for the prediction of type 2 diabetes are one example (131). Another example is the *ATM* locus for metformin treatment response, the most widely replicated diabetes pharmacogenomic finding; each copy of the variant allele was associated with only a 0.1% greater reduction in HbA_{1c} (81). Even if patients had whole-genome sequence data already available in their medical records, this treatment effect is still too small to be clinically useful. For MODY, which often has a clear genetic cause, most cases appear to be undiagnosed, in part because of uncertainty from clinicians about the clinical benefits of genetic testing (132). Making treatment decisions based on genetic tests that lack clinical validity, clinical utility, or both can have unintended consequences, including withholding beneficial treatments, an unnecessary increase in costs, or the use of drugs with harmful effects.

For examples of actionable genetic tests that meet the criteria for clinical validity and clinical utility, areas of medicine other than diabetes are instructive. Abacavir is a first-line treatment for HIV that is effective and well tolerated in most people, but about 5% of persons exposed to this drug develop a serious hypersensitivity reaction characterized by rash, fever, and damage to multiple organ systems (133). Initial studies of men of mostly European ancestry identified the cause of this ADR to be the *HLA-B*57:01* variant (91,134), with an OR of well over 100 (119), and this finding was later extended to other populations (135). A double-blind RCT demonstrated that genetic testing for the *HLA-B*57:01* variant prevented all cases of abacavir-induced hypersensitivity (136), and genetic testing is now routinely conducted before initiating treatment with this drug.

For another severe immune-mediated ADR, the translation of genetic testing into practice followed a different course. Stevens-Johnson syndrome/toxic epidermal necrolysis (SJS/TEN) related to carbamazepine use was found to be strongly associated with the *HLA* variant *HLA-B*15:02* in several Asian populations (92,137), and screening for this variant can effectively prevent the ADR (137). In 2008, Hong Kong instituted a policy of routine genetic screening before initiating carbamazepine treatment, but providers simply avoided prescribing this medication and instead prescribed other antiepileptic drugs that can also cause

Table 3—ACCE criteria for the implementation of a genetic test

	Key questions	Example applications to diabetes
Analytic validity	How often does the genetic test fail to give a useable result?	For a GRS for type 2 diabetes risk, how often does the test result in a genotype at each locus?
	What is the sensitivity and specificity of the test for a genetic variant?	For the rare *HNF1A* missense variant associated with diabetes risk, how often does the test detect the variant when it is present and how often does the test result in a false positive when the variant is not present?
	What is the within- and between-laboratory precision?	Are standardized methods used in different laboratories? How does the accuracy of the test vary from laboratory to laboratory?
	Is confirmatory testing required?	When a pathogenic variant in a MODY gene is identified from high-throughput sequencing, is confirmatory genotyping by another method required?
Clinical validity	What is the quality of the disease phenotype or treatment response measurement? What is the quality of the study designs used to evaluate these outcomes?	For a GRS that discriminates type 1 from type 2 diabetes, are standardized definitions of diabetes type used? Was the study design cross-sectional or longitudinal?
	What is the prevalence of the phenotype or the distribution of treatment response in the studied populations?	For a pharmacogenomic test for metformin treatment response, how is treatment failure/success defined and what is the rate of treatment failure/success in the studied population?
	What is the sensitivity and specificity of the test for the disease phenotype or treatment response? What is the magnitude and precision of the genotype-phenotype relationship?	What is the relative change in HbA_{1c} lowering associated with the *ATM* locus for metformin treatment response?
	What are the genetic or environmental modifiers of the genotype-phenotype relationship?	Do diabetes severity, obesity, the use of additional drug therapies, or other factors modify the pharmacogenomic associations for metformin treatment response?
	Has the test been adequately validated on all populations in which it may be offered?	For a GRS for type 2 diabetes risk, has the test been validated among obese individuals and among persons from different racial/ethnic populations?
Clinical utility	What is the impact of a positive or negative test on patient care in terms of health outcomes?	What is the absolute change in HbA_{1c} lowering associated with the *ATM* locus for metformin treatment response? Does this pharmacogenomic test result in projected or tangible health benefits in terms of clinical complications, such as a reduced risk of blindness or cardiovascular disease?
	What are the financial costs associated with testing and the economic benefits associated with actions resulting from testing?	What is the cost to test for the common *TCF7L2* variant for type 2 diabetes risk, including laboratory, reporting, educational, and counseling costs? What is the cost per case of diabetes predicted? Does this test result in cost-effective screening or prevention of diabetes later in life?
	What educational materials for patients have been developed and validated?	Do educational materials clearly explain the magnitude of the increased risk of type 2 diabetes associated with the common *TCF7L2* variant in relation to other known risk factors?
Ethical, legal, and social implications	What is known about how this test could lead to stigmatization, discrimination, privacy/confidentiality, and personal/family social issues?	Does the identification of a pathogenic variant responsible for a hereditary form of early-onset diabetes have implications for family planning?
	Are there legal issues regarding consent, ownership of data and/or samples, patents, licensing, proprietary testing, disclosure, or reporting requirements?	Are physicians providing sufficient information about the potential risks and benefits from genetic testing so that patients can make informed decisions?
	What safeguards have been described and are these safeguards effective?	

Adapted from refs. 126,127.

severe skin reactions. As a result, the overall rate of SJS/TEN in Hong Kong did not change (138). This unintended result of a policy decision on a genetic test with clinical validity and clinical utility demonstrates that other concerns, such as convenience or cost, may influence the harms or benefits for a population.

Even for well-replicated pharmacogenomic findings with large effects, there can be disagreement about what constitutes clinical utility. Warfarin is an anticoagulant drug that is notoriously difficult to dose because of wide interindividual variability in the anticoagulant effect, much of which is explained by variation in the pharmacokinetic genes *CYP2C9* and *VKORC1*. The benefits of testing for variants in these genes have even been evaluated in three large well-designed RCTs, two of which demonstrated that genetic testing resulted in a shorter time to therapeutic effect but no reduction in thrombotic or bleeding complications (139–141). Opinions differ about whether pharmacogenomic testing has a role in warfarin treatment (142–144), and there has been a broader debate about whether RCTs are required to demonstrate the utility of a genetic test (79,145,146). Nonetheless, genetic testing for warfarin treatment has been minimal in practice. In the absence of consistent recommendations about when to use genetic information in prescribing and other treatment decisions, health care providers, patients, and also payers and health care systems will have an important voice in this discussion.

ADDITIONAL BARRIERS TO THE IMPLEMENTATION OF GENOMICS

The implementation of genomics in practice has generally involved a decision about whether to order a genetic test before starting a new treatment. For patients who have already undergone preemptive genotyping or sequencing, which is not yet common but likely to increase in the future (147–149), the decision may be about whether to use a patient's existing genetic information to guide treatment. With either approach, the main barrier to the implementation of genomics for diabetes care at this time is a robust lack of actionable findings in the areas of diabetes and cardiovascular medicine. There are also practical issues that must be addressed, which include the storage and integration of complex genomic data into the electronic health record, the interpretation of these data in an accessible format, and the creation of clinical decision support (CDS) tools that aid health care providers in using this information to make treatment decisions at the point of care (150). Various groups have made important contributions in developing model systems that address these problems (147,148,151,152).

CDS tools, which have been in use for many years for diabetes management, provide automated testing or treatment recommendations based on information in the electronic health record (153). The CDS tools for genomics that are under development rely on externally curated data resources, such as PharmGKB (154) and ClinVar (155), to determine which potentially actionable variants to incorporate and how to translate genotypes into expected phenotypes. PharmGKB currently includes information about pharmacogenomic associations for metformin and sulfonylureas; however, none of the associations meet the criteria for the highest level of evidence, and there is no recommendation for pharmacogenomic testing for these drugs (156). In drug-specific evidence summaries and guidelines, the Clinical Pharmacogenetics Implementation Consortium has begun to provide standardized terms, sample text for electronic health record documentation and point of care alerts, and clinical implementation workflows, which will be a valuable resource as the role of pharmacogenomics expands in clinical practice (157,158).

The increasing availability of CDS tools and online databases with information on actionable genomic findings will facilitate the use of genomic testing for diabetes by clinicians. However, because most clinicians are not yet comfortable using genomic information to make clinical decisions (159,160), improved education for both trainees and practicing clinicians is needed, and opportunities for both are improving (161,162). Professional societies and specialty boards will play an important role in the integration of genomics in physician training and continuing medical education; credentialing requirements for ordering genomic tests could help to ensure that such tests are used appropriately (163). If the use of genomic information ultimately proves useful in the care of patients with diabetes, even with adequate educational and training opportunities for clinicians, financial incentives may be required to spur widespread implementation, as was the case with the increased adoption of electronic health records after legislation was passed that provided reimbursements to hospitals and providers for this purpose (9–11).

CONCLUSIONS

At this time, there are few if any actionable genomic findings for diabetes that are ready for implementation. However, the increasing availability of genomic data in large populations linked with electronic health data may become a powerful resource for genomic discovery, and examples from other areas of medicine offer lessons about the limitations of these data that can help guide the direction of future research. Whether genomic information should be used in clinical practice requires a framework for evaluating the validity and clinical utility of this approach, an improved integration of genomic data into electronic health records, and the clinical decision support and educational resources for clinicians to use these data. Efforts to identify optimal approaches in all of these domains are creating a growing body of evidence that may help to bring diabetes into the era of genomic medicine.

Funding. J.S.F. was supported by National Heart, Lung, and Blood Institute grant K08HL116640.

Funding agencies did not influence the design and conduct of the study, collection, management, analysis, and interpretation of the data or preparation, review, or approval of the manuscript.

Duality of Interest. B.M.P. serves on the Data and Safety Monitoring Board of a clinical trial of a device funded by the manufacturer (Zoll LifeCor) and on the Steering Committee of the Yale University Open Data Access Project funded by Johnson & Johnson. No other potential conflicts of interest relevant to this article were reported.

References

1. Cohen J, Pertsemlidis A, Kotowski IK, Graham R, Garcia CK, Hobbs HH. Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in PCSK9. Nat Genet 2005;37:161–165
2. Cohen JC, Boerwinkle E, Mosley TH Jr, Hobbs HH. Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. N Engl J Med 2006;354:1264–1272
3. Robinson JG, Farnier M, Krempf M, et al. Efficacy and safety of alirocumab in reducing lipids and cardiovascular events. N Engl J Med 2015;372:1489–1499
4. Sabatine MS, Giugliano RP, Wiviott SD, et al.; Open-Label Study of Long-Term Evaluation against LDL Cholesterol (OSLER) Investigators. Efficacy and safety of evolocumab in reducing lipids and cardiovascular events. N Engl J Med 2015;372:1500–1509
5. McLeod HL. Cancer pharmacogenomics: early promise, but concerted effort needed. Science 2013;339:1563–1566
6. Chapman PB, Hauschild A, Robert C, et al.; BRIM-3 Study Group. Improved survival with vemurafenib in melanoma with BRAF V600E mutation. N Engl J Med 2011;364:2507–2516
7. Jameson JL, Longo DL. Precision medicine–personalized, problematic, and promising. N Engl J Med 2015;372:2229–2234
8. Hayden EC. Technology: The $1,000 genome. Nature 2014;507:294–295
9. Adler-Milstein J, Furukawa MF, King J, Jha AK. Early results from the hospital Electronic Health Record Incentive Programs. Am J Manag Care 2013;19:e273–e284
10. Xierali IM, Hsiao CJ, Puffer JC, et al. The rise of electronic health record adoption among

family physicians. Ann Fam Med 2013;11: 14–19
11. Lehmann CU, O'Connor KG, Shorte VA, Johnson TD. Use of electronic health record systems by office-based pediatricians. Pediatrics 2015;135:e7–e15
12. McCarthy MI. Genomic medicine at the heart of diabetes management. Diabetologia 2015;58:1725–1729
13. Geiss LS, Wang J, Cheng YJ, et al. Prevalence and incidence trends for diagnosed diabetes among adults aged 20 to 79 years, United States, 1980-2012. JAMA 2014;312:1218–1226
14. Centers for Disease Control and Prevention. *National Diabetes Statistics Report, 2014*. National Center for Chronic Disease Prevention and Health Promotion, Atlanta, GA, 2014
15. Morris AP, Voight BF, Teslovich TM, et al.; Wellcome Trust Case Control Consortium; Meta-Analyses of Glucose and Insulin-related traits Consortium (MAGIC) Investigators; Genetic Investigation of ANthropometric Traits (GIANT) Consortium; Asian Genetic Epidemiology Network–Type 2 Diabetes (AGEN-T2D) Consortium; South Asian Type 2 Diabetes (SAT2D) Consortium; DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nat Genet 2012;44:981–990
16. Scott RA, Lagou V, Welch RP, et al. Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. Nat Genet 2012;44:991–1005
17. DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium; Asian Genetic Epidemiology Network Type 2 Diabetes (AGEN-T2D) Consortium; South Asian Type 2 Diabetes (SAT2D) Consortium; Mexican American Type 2 Diabetes (MAT2D) Consortium; Type 2 Diabetes Genetic Exploration by Nex-generation sequencing in muylti-Ethnic Samples (T2D-GENES) Consortium. Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. Nat Genet 2014;46:234–244
18. Sladek R, Prokopenko I. Genome-Wide Association Studies of Type 2 Diabetes. In *The Genetics of Type 2 Diabetes and Related Traits*. Florez JC, Ed. Springer International Publishing, 2016, p. 13–61
19. Zeggini E, Scott LJ, Saxena R, et al. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. Nat Genet 2008;40:638–645
20. Estrada K, Aukrust I, Bjorkhaug L, et al. Association of a low-frequency variant in HNF1A with type 2 diabetes in a Latino population. JAMA 2014;311:2305–2314
21. Moltke I, Grarup N, Jorgensen ME, et al. A common Greenlandic TBC1D4 variant confers muscle insulin resistance and type 2 diabetes. Nature 2014;512:190–193
22. Fuchsberger C, Flannick J, Teslovich TM, et al. The genetic architecture of type 2 diabetes. Nature 2016;536:41–47
23. Meigs JB, Shrader P, Sullivan LM, et al. Genotype score in addition to common risk factors for prediction of type 2 diabetes. N Engl J Med 2008;359:2208–2219
24. Vassy JL, Hivert MF, Porneala B, et al. Polygenic type 2 diabetes prediction at the limit of common variant detection. Diabetes 2014;63: 2172–2182
25. Pepe MS, Janes H, Longton G, Leisenring W, Newcomb P. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. Am J Epidemiol 2004;159:882–890
26. Lyssenko V, Laakso M. Genetic screening for the risk of type 2 diabetes: worthless or valuable? Diabetes Care 2013;36(Suppl. 2):S120–S126
27. Yuan W, Xia Y, Bell CG, et al. An integrated epigenomic analysis for type 2 diabetes susceptibility loci in monozygotic twins. Nat Commun 2014;5:5719
28. Chambers JC, Loh M, Lehne B, et al. Epigenome-wide association of DNA methylation markers in peripheral blood from Indian Asians and Europeans with incident type 2 diabetes: a nested case-control study. Lancet Diabetes Endocrinol 2015;3:526–534
29. Wang TJ, Larson MG, Vasan RS, et al. Metabolite profiles and the risk of developing diabetes. Nat Med 2011;17:448–453
30. Walford GA, Porneala BC, Dauriz M, et al. Metabolite traits and genetic risk provide complementary information for the prediction of future type 2 diabetes. Diabetes Care 2014;37: 2508–2514
31. Knowler WC, Barrett-Connor E, Fowler SE, et al.; Diabetes Prevention Program Research Group. Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. N Engl J Med 2002;346:393–403
32. Florez JC, Jablonski KA, Bayley N, et al.; Diabetes Prevention Program Research Group. TCF7L2 polymorphisms and progression to diabetes in the Diabetes Prevention Program. N Engl J Med 2006;355:241–250
33. Hivert MF, Jablonski KA, Perreault L, et al.; DIAGRAM Consortium; Diabetes Prevention Program Research Group. Updated genetic score based on 34 confirmed type 2 diabetes Loci is associated with diabetes incidence and regression to normoglycemia in the diabetes prevention program. Diabetes 2011;60:1340–1348
34. Hollands GJ, French DP, Griffin SJ, et al. The impact of communicating genetic risks of disease on risk-reducing health behaviour: systematic review with meta-analysis. BMJ 2016;352: i1102
35. Grant RW, O'Brien KE, Waxler JL, et al. Personalized genetic risk counseling to motivate diabetes prevention: a randomized trial. Diabetes Care 2013;36:13–19
36. Ogden CL, Carroll MD, Kit BK, Flegal KM. Prevalence of childhood and adult obesity in the United States, 2011-2012. JAMA 2014;311: 806–814
37. Farmer A, Fox R. Diagnosis, classification, and treatment of diabetes. BMJ 2011;342: d3319
38. Oram RA, Patel K, Hill A, et al. A type 1 diabetes genetic risk score can aid discrimination between type 1 and type 2 diabetes in young adults. Diabetes Care 2016;39:337–344
39. Pearson ER, Starkey BJ, Powell RJ, Gribble FM, Clark PM, Hattersley AT. Genetic cause of hyperglycaemia and response to treatment in diabetes. Lancet 2003;362:1275–1281
40. Pearson ER, Flechtner I, Njolstad PR, et al. Switching from insulin to oral sulfonylureas in patients with diabetes due to Kir6.2 mutations. N Engl J Med 2006;355:467–477
41. Fajans SS, Bell GI. MODY: history, genetics, pathophysiology, and clinical decision making. Diabetes Care 2011;34:1878–1884
42. Vaxillaire M, Froguel P. Monogenic diabetes in the young, pharmacogenetics and relevance to multifactorial forms of type 2 diabetes. Endocr Rev 2008;29:254–264
43. Gregg EW, Li Y, Wang J, et al. Changes in diabetes-related complications in the United States, 1990-2010. N Engl J Med 2014;370:1514–1523
44. The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. The Diabetes Control and Complications Trial Research Group. N Engl J Med 1993;329:977–986
45. Nathan DM, Cleary PA, Backlund JY, et al.; Diabetes Control and Complications Trial/ Epidemiology ofDiabetes Interventions and Complications (DCCT/EDIC) Study Research Group. Intensive diabetes treatment and cardiovascular disease in patients with type 1 diabetes. N Engl J Med 2005;353:2643–2653
46. Zoungas S, Chalmers J, Neal B, et al.; ADVANCE-ON Collaborative Group. Follow-up of blood-pressure lowering and glucose control in type 2 diabetes. N Engl J Med 2014;371:1392–1406
47. Holman RR, Paul SK, Bethel MA, Matthews DR, Neil HA. 10-year follow-up of intensive glucose control in type 2 diabetes. N Engl J Med 2008;359:1577–1589
48. Hayward RA, Reaven PD, Wiitala WL, et al. Follow-up of glycemic control and cardiovascular outcomes in type 2 diabetes. N Engl J Med 2015;372:2197–2206
49. Nathan DM. Diabetes: advances in diagnosis and treatment. JAMA 2015;314:1052–1062
50. Lipska KJ, Krumholz H, Soones T, Lee SJ. Polypharmacy in the aging patient: a review of glycemic control in older adults with type 2 diabetes. JAMA 2016;315:1034–1045
51. Holman RR, Sourij H, Califf RM. Cardiovascular outcome trials of glucose-lowering drugs or strategies in type 2 diabetes. Lancet 2014; 383:2008–2017
52. American Diabetes Association. Summary of Revisions. In Standards of Medical Care in Diabetes—2016. Diabetes Care 2016;39(Suppl. 1):S4–S5
53. Effect of intensive blood-glucose control with metformin on complications in overweight patients with type 2 diabetes (UKPDS 34). UK Prospective Diabetes Study (UKPDS) Group. Lancet 1998;352:854–865
54. Hong J, Zhang Y, Lai S, et al. Effects of metformin versus glipizide on cardiovascular outcomes in patients with type 2 diabetes and coronary artery disease. Diabetes Care 2013; 36:1304–1311
55. Walford GA, Colomo N, Todd JN, et al. The study to understand the genetics of the acute response to metformin and glipizide in humans (SUGAR-MGH): design of a pharmacogenetic resource for type 2 diabetes. PLoS One 2015;10: e0121553
56. Zeitler P, Hirst K, Pyle L, et al. A clinical trial to maintain glycemic control in youth with

type 2 diabetes. N Engl J Med 2012;366:2247–2256

57. Kahn SE, Haffner SM, Heise MA, et al. Glycemic durability of rosiglitazone, metformin, or glyburide monotherapy. N Engl J Med 2006;355:2427–2443

58. Cook MN, Girman CJ, Stein PP, Alexander CM. Initial monotherapy with either metformin or sulphonylureas often fails to achieve or maintain current glycaemic goals in patients with type 2 diabetes in UK primary care. Diabet Med 2007;24:350–358

59. Donnelly LA, Doney AS, Hattersley AT, Morris AD, Pearson ER. The effect of obesity on glycaemic response to metformin or sulphonylureas in type 2 diabetes. Diabet Med 2006;23:128–133

60. Pirmohamed M. Personalized pharmacogenomics: predicting efficacy and adverse drug reactions. Annu Rev Genomics Hum Genet 2014;15:349–370

61. Manolopoulos VG, Ragia G, Tavridou A. Pharmacogenomics of oral antidiabetic medications: current data and pharmacoepigenomic perspective. Pharmacogenomics 2011;12:1161–1191

62. Becker ML, Pearson ER, Tkac I. Pharmacogenetics of oral antidiabetic drugs. Int J Endocrinol 2013;2013:686315

63. Pollastro C, Ziviello C, Costa V, Ciccodicola A. Pharmacogenomics of drug response in type 2 diabetes: toward the definition of tailored therapies? PPAR Res 2015;2015:415149

64. Zhou K, Donnelly L, Burch L, et al. Loss-of-function CYP2C9 variants improve therapeutic response to sulfonylureas in type 2 diabetes: a Go-DARTS study. Clin Pharmacol Ther 2010;87:52–56

65. Holstein A, Plaschke A, Ptak M, et al. Association between CYP2C9 slow metabolizer genotypes and severe hypoglycaemia on medication with sulphonylurea hypoglycaemic agents. Br J Clin Pharmacol 2005;60:103–106

66. Feng Y, Mao G, Ren X, et al. Ser1369Ala variant in sulfonylurea receptor gene ABCC8 is associated with antidiabetic efficacy of gliclazide in Chinese type 2 diabetic patients. Diabetes Care 2008;31:1939–1944

67. Javorsky M, Klimcakova L, Schroner Z, et al. KCNJ11 gene E23K variant and therapeutic response to sulfonylureas. Eur J Intern Med 2012;23:245–249

68. Sesti G, Laratta E, Cardellini M, et al. The E23K variant of KCNJ11 encoding the pancreatic beta-cell adenosine 5′-triphosphate-sensitive potassium channel subunit Kir6.2 is associated with an increased risk of secondary failure to sulfonylurea in patients with type 2 diabetes. J Clin Endocrinol Metab 2006;91:2334–2339

69. Nikolac N, Simundic AM, Katalinic D, Topic E, Cipak A, Zjacic Rotkvic V. Metabolic control in type 2 diabetes is associated with sulfonylurea receptor-1 (SUR-1) but not with KCNJ11 polymorphisms. Arch Med Res 2009;40:387–392

70. Holstein A, Hahn M, Stumvoll M, Kovacs P. The E23K variant of KCNJ11 and the risk for severe sulfonylurea-induced hypoglycemia in patients with type 2 diabetes. Horm Metab Res 2009;41:387–390

71. Graham GG, Punt J, Arora M, et al. Clinical pharmacokinetics of metformin. Clin Pharmacokinet 2011;50:81–98

72. Pawlyk AC, Giacomini KM, McKeon C, Shuldiner AR, Florez JC. Metformin pharmacogenomics: current status and future directions. Diabetes 2014;63:2590–2599

73. Dujic T, Zhou K, Donnelly LA, Tavendale R, Palmer CN, Pearson ER. Association of organic cation transporter 1 with intolerance to metformin in type 2 diabetes: a GoDARTS study. Diabetes 2015;64:1786–1793

74. Cho HJ, Lee SY, Kim YG, et al. Effect of genetic polymorphisms on the pharmacokinetics and efficacy of glimepiride in a Korean population. Clin Chim Acta 2011;412:1831–1834

75. Becker ML, Visser LE, Trienekens PH, Hofman A, van Schaik RH, Stricker BH. Cytochrome P450 2C9 *2 and *3 polymorphisms and the dose and effect of sulfonylurea in type II diabetes mellitus. Clin Pharmacol Ther 2008;83:288–292

76. Zhou K, Donnelly LA, Kimber CH, et al. Reduced-function SLC22A1 polymorphisms encoding organic cation transporter 1 and glycemic response to metformin: a GoDARTS study. Diabetes 2009;58:1434–1439

77. Christensen MM, Brasch-Andersen C, Green H, et al. The pharmacogenetics of metformin and its impact on plasma metformin steady-state levels and glycosylated hemoglobin A1c. Pharmacogenet Genomics 2011;21:837–850

78. Ioannidis JP, Tarone R, McLaughlin JK. The false-positive to false-negative ratio in epidemiologic studies. Epidemiology 2011;22:450–456

79. Aslibekyan S, Claas SA, Arnett DK. To replicate or not to replicate: the case of pharmacogenetic studies: Establishing validity of pharmacogenomic findings: from replication to triangulation. Circ Cardiovasc Genet 2013;6:409–412; discussion 412

80. GoDARTS and UKPDS Diabetes Pharmacogenetics Study Group, Wellcome Trust Case Control Consortium, Zhou K, et al. Common variants near ATM are associated with glycemic response to metformin in type 2 diabetes. Nat Genet 2011;43:117–120

81. van Leeuwen N, Nijpels G, Becker ML, et al. A gene variant near ATM is significantly associated with metformin treatment response in type 2 diabetes: a replication and meta-analysis of five cohorts. Diabetologia 2012;55:1971–1977

82. Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. Nature 2009;461:747–753

83. Kryukov GV, Pennacchio LA, Sunyaev SR. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. Am J Hum Genet 2007;80:727–739

84. Pritchard JK, Cox NJ. The allelic architecture of human disease genes: common disease-common variant...or not? Hum Mol Genet 2002;11:2417–2423

85. Tucker T, Marra M, Friedman JM. Massively parallel sequencing: the next big thing in genetic medicine. Am J Hum Genet 2009;85:142–154

86. Nelson MR, Wegmann D, Ehm MG, et al. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science 2012;337:100–104

87. Gordon AS, Tabor HK, Johnson AD, et al. Quantifying rare, deleterious variation in 12 human cytochrome P450 drug-metabolism genes in a large-scale exome dataset. Hum Mol Genet 2014;23:1957–1963

88. Burton PR, Hansell AL, Fortier I, et al. Size matters: just how big is BIG?: Quantifying realistic sample size requirements for human genome epidemiology. Int J Epidemiol 2009;38:263–273

89. Motsinger-Reif AA, Jorgenson E, Relling MV, et al. Genome-wide association studies in pharmacogenomics: successes and lessons. Pharmacogenet Genomics 2013;23:383–394

90. Daly AK, Donaldson PT, Bhatnagar P, et al. HLA-B*5701 genotype is a major determinant of drug-induced liver injury due to flucloxacillin. Nat Genet 2009;41:816–819

91. Mallal S, Nolan D, Witt C, et al. Association between presence of HLA-B*5701, HLA-DR7, and HLA-DQ3 and hypersensitivity to HIV-1 reverse-transcriptase inhibitor abacavir. Lancet 2002;359:727–732

92. Chung WH, Hung SI, Hong HS, et al. Medical genetics: a marker for Stevens-Johnson syndrome. Nature 2004;428:486

93. Bis JC, Sitlani C, Irvin R, et al. Drug-gene interactions of antihypertensive medications and risk of incident cardiovascular disease: a pharmacogenomics study from the CHARGE Consortium. PLoS One 2015;10:e0140496

94. Avery CL, Sitlani CM, Arking DE, et al. Drug-gene interactions and the search for missing heritability: a cross-sectional pharmacogenomics study of the QT interval. Pharmacogenomics J 2014;14:6–13

95. Postmus I, Trompet S, Deshmukh HA, et al. Pharmacogenetic meta-analysis of genome-wide association studies of LDL cholesterol response to statins. Nat Commun 2014;5:5068

96. Lonsdale J, Thomas J, Salvatore M, et al. The Genotype-Tissue Expression (GTEx) project. Nat Genet 2013;45:580–585

97. Koester SE, Insel TR. Understanding how noncoding genomic polymorphisms affect gene expression. Mol Psychiatry 2016;21:448–449

98. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature 2012;489:57–74

99. Asimit J, Zeggini E. Rare variant association analysis methods for complex traits. Annu Rev Genet 2010;44:293–308

100. Liu DJ, Peloso GM, Zhan X, et al. Meta-analysis of gene-level tests for rare variant association. Nat Genet 2014;46:200–204

101. Lange LA, Hu Y, Zhang H, et al. Whole-exome sequencing identifies rare and low-frequency coding variants associated with LDL cholesterol. Am J Hum Genet 2014;94:233–245

102. Psaty BM, O'Donnell CJ, Gudnason V, et al.; CHARGE Consortium. Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium: design of prospective meta-analyses of genome-wide association studies from 5 cohorts. Circ Cardiovasc Genet 2009;2:73–80

103. The German National Cohort. Aims, study design and organization. Eur J Epidemiol 2014;29:371–382

104. Sudlow C, Gallacher J, Allen N, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med 2015;12:e1001779

105. Gaziano JM, Concato J, Brophy M, et al. Million Veteran Program: a mega-biobank to study genetic influences on health and disease. J Clin Epidemiol 2016;70:214–223
106. Song Y, Skinner J, Bynum J, Sutherland J, Wennberg JE, Fisher ES. Regional variations in diagnostic practices. N Engl J Med 2010;363:45–53
107. Lindenauer PK, Lagu T, Shieh MS, Pekow PS, Rothberg MB. Association of diagnostic coding with trends in hospitalizations and mortality of patients with pneumonia, 2003-2009. JAMA 2012;307:1405–1413
108. Psaty BM, Boineau R, Kuller LH, Luepker RV. The potential costs of upcoding for heart failure in the United States. Am J Cardiol 1999; 84:108–109, A9
109. Madigan D, Ryan PB, Schuemie M, et al. Evaluating the impact of database heterogeneity on observational study results. Am J Epidemiol 2013;178:645–651
110. Ginde AA, Blanc PG, Lieberman RM, Camargo CA Jr. Validation of ICD-9-CM coding algorithm for improved identification of hypoglycemia visits. BMC Endocr Disord 2008;8:4
111. Schelleman H, Bilker WB, Brensinger CM, Wan F, Hennessy S. Anti-infectives and the risk of severe hypoglycemia in users of glipizide or glyburide. Clin Pharmacol Ther 2010;88:214–222
112. Cutrona SL, Toh S, Iyer A, et al. Validation of acute myocardial infarction in the Food and Drug Administration's Mini-Sentinel program. Pharmacoepidemiol Drug Saf 2013;22:40–54
113. Floyd JS, Blondon M, Moore KP, Boyko EJ, Smith NL. Validation of methods for assessing cardiovascular disease using electronic health data in a cohort of Veterans with diabetes. Pharmacoepidemiol Drug Saf 2016;25:467–471
114. Psaty BM, Delaney JA, Arnold AM, et al. Study of cardiovascular health outcomes in the era of claims data: the Cardiovascular Health Study. Circulation 2016;133:156–164
115. Agarwal SK, Wruck L, Quibrera M, et al. Temporal trends in hospitalization for acute decompensated heart failure in the United States, 1998-2011. Am J Epidemiol 2016;183:462–470
116. Chubak J, Pocobelli G, Weiss NS. Tradeoffs between accuracy measures for electronic health care data algorithms. J Clin Epidemiol 2012;65:343–349.e2
117. Link E, Parish S, Armitage J, et al. SLCO1B1 variants and statin-induced myopathy–a genomewide study. N Engl J Med 2008;359:789–799
118. Floyd JS, Heckbert SR, Weiss NS, Carrell DS, Psaty BM. Use of administrative data to estimate the incidence of statin-related rhabdomyolysis. JAMA 2012;307:1580–1582
119. Sousa-Pinto B, Pinto-Ramos J, Correia C, et al. Pharmacogenetics of abacavir hypersensitivity: a systematic review and meta-analysis of the association with HLA-B*57:01. J Allergy Clin Immunol 2015;136:1092–1094.e3
120. Carr DF, O'Meara H, Jorgensen AL, et al. SLCO1B1 genetic variant associated with statin-induced myopathy: a proof-of-concept study using the clinical practice research datalink. Clin Pharmacol Ther 2013;94:695–701
121. Manuel DG, Rosella LC, Stukel TA. Importance of accurately identifying disease in studies using electronic health records. BMJ 2010;341:c4226
122. Ramsey BW, Davies J, McElvaney NG, et al. A CFTR potentiator in patients with cystic fibrosis and the G551D mutation. N Engl J Med 2011;365: 1663–1672
123. U.S. Food and Drug Administration. Table of pharmacogenomic biomarkers in drug labeling [Internet]. Available from http://www.fda.gov/drugs/scienceresearch/researchareas/pharmacogenetics/ucm083378.htm. Accessed 15 August 2016
124. Sharfstein J. FDA regulation of laboratory-developed diagnostic tests: protect the public, advance the science. JAMA 2015;313:667–668
125. Wang B, Canestaro WJ, Choudhry NK. Clinical evidence supporting pharmacogenomic biomarker testing provided in US Food and Drug Administration drug labels. JAMA Intern Med 2014;174:1938–1944
126. Centers for Disease Control and Prevention. ACCE model process for evaluating genetic tests [Internet]. Available from http://www.cdc.gov/genomics/gtesting/ACCE/. Acessed 26 March 2016
127. Teutsch SM, Bradley LA, Palomaki GE, et al. The Evaluation of Genomic Applications in Practice and Prevention (EGAPP) Initiative: methods of the EGAPP Working Group. Genet Med 2009;11:3–14
128. Richards S, Aziz N, Bale S, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med 2015;17:405–424
129. Manrai AK, Ioannidis JP, Kohane IS. Clinical genomics: from pathogenicity claims to quantitative risk estimates. JAMA 2016;315:1233–1234
130. Flannick J, Beer NL, Bick AG, et al. Assessing the phenotypic effects in the general population of rare variants in genes for a dominant Mendelian form of diabetes. Nat Genet 2013; 45:1380–1385
131. Palomaki GE, Melillo S, Marrone M, Douglas MP. Use of genomic panels to determine risk of developing type 2 diabetes in the general population: a targeted evidence-based review. Genet Med 2013;15:600–611
132. van der Zwaag AM, Weinreich SS, Bosma AR, et al. Current and best practices of genetic testing for maturity onset diabetes of the young: views of professional experts. Public Health Genomics 2015;18:52–59
133. Gunthard HF, Aberg JA, Eron JJ, et al.; International Antiviral Society-USA Panel. Antiretroviral treatment of adult HIV infection: 2014 recommendations of the International Antiviral Society-USA Panel. JAMA 2014;312:410–425
134. Hetherington S, Hughes AR, Mosteller M, et al. Genetic variations in HLA-B region and hypersensitivity reactions to abacavir. Lancet 2002;359:1121–1122
135. Saag M, Balu R, Phillips E, et al.; Study of Hypersensitivity to Abacavir and Pharmacogenetic Evaluation Study Team. High sensitivity of human leukocyte antigen-b*5701 as a marker for immunologically confirmed abacavir hypersensitivity in white and black patients. Clin Infect Dis 2008;46:1111–1118
136. Mallal S, Phillips E, Carosi G, et al.; PREDICT-1 Study Team. HLA-B*5701 screening for hypersensitivity to abacavir. N Engl J Med 2008;358:568–579
137. Chen P, Lin JJ, Lu CS, et al. Carbamazepine-induced toxic effects and HLA-B*1502 screening in Taiwan. N Engl J Med 2011;364:1126–1133
138. Chen Z, Liew D, Kwan P. Effects of a HLA-B*15:02 screening policy on antiepileptic drug use and severe skin reactions. Neurology 2014;83:2077–2084
139. Kimmel SE, French B, Kasner SE, et al. A pharmacogenetic versus a clinical algorithm for warfarin dosing. N Engl J Med 2013;369:2283–2293
140. Pirmohamed M, Burnside G, Eriksson N, et al. Group E-P: A randomized trial of genotype-guided dosing of warfarin. N Engl J Med 2013;369: 2294–2303
141. Verhoef TI, Ragia G, de Boer A, et al.; EU-PACT Group. A randomized trial of genotype-guided dosing of acenocoumarol and phenprocoumon. N Engl J Med 2013;369:2304–2312
142. Furie B. Do pharmacogenetics have a role in the dosing of vitamin K antagonists? N Engl J Med 2013;369:2345–2346
143. Pirmohamed M, Kamali F, Daly AK, Wadelius M. Oral anticoagulation: a critique of recent advances and controversies. Trends Pharmacol Sci 2015;36:153–163
144. Cavallari LH, Nutescu EA. Warfarin pharmacogenetics: to genotype or not to genotype, that is the question. Clin Pharmacol Ther 2014; 96:22–24
145. Ioannidis JP, Khoury MJ. Are randomized trials obsolete or more important than ever in the genomic era? Genome Med 2013;5:32
146. Gillis NK, Innocenti F. Evidence required to demonstrate clinical utility of pharmacogenetic testing: the debate continues. Clin Pharmacol Ther 2014;96:655–657
147. Gottesman O, Scott SA, Ellis SB, et al. The CLIPMERGE PGx Program: clinical implementation of personalized medicine through electronic health records and genomics-pharmacogenomics. Clin Pharmacol Ther 2013; 94:214–217
148. Bielinski SJ, Olson JE, Pathak J, et al. Preemptive genotyping for personalized medicine: design of the right drug, right dose, right time-using genomic data to individualize treatment protocol. Mayo Clin Proc 2014;89:25–33
149. Rasmussen-Torvik LJ, Stallings SC, Gordon AS, et al. Design and anticipated outcomes of the eMERGE-PGx project: a multicenter pilot for preemptive pharmacogenomics in electronic health record systems. Clin Pharmacol Ther 2014;96:482–489
150. Relling MV, Evans WE. Pharmacogenomics in the clinic. Nature 2015;526:343–350
151. Gottesman O, Kuivaniemi H, Tromp G, et al. The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. Genet Med 2013;15:761–771
152. Weitzel KW, Alexander M, Bernhardt BA, et al. The IGNITE network: a model for genomic medicine implementation and research. BMC Med Genomics 2016;9:1
153. Jeffery R, Iserman E, Haynes RB; CDSS Systematic Review Team. Can computerized clinical decision support systems improve diabetes management? A systematic review and meta-analysis. Diabet Med 2013;30:739–745
154. Thorn CF, Klein TE, Altman RB. Pharmacogenomics and bioinformatics: PharmGKB. Pharmacogenomics 2010;11:501–505
155. Rehm HL, Berg JS, Brooks LD, et al.; ClinGen. ClinGen–the Clinical Genome Resource. N Engl J Med 2015;372:2235–2242

156. PharmGKB. Available from https://www.pharmgkb.org. Accessed 12 July 2016
157. Martin MA, Hoffman JM, Freimuth RR, et al.; Clinical Pharmacogenetics Implementation Consortium. Clinical pharmacogenetics implementation consortium guidelines for HLA-B genotype and abacavir dosing: 2014 update. Clin Pharmacol Ther 2014;95:499–500
158. Ramsey LB, Johnson SG, Caudle KE, et al. The clinical pharmacogenetics implementation consortium guideline for SLCO1B1 and simvastatin-induced myopathy: 2014 update. Clin Pharmacol Ther 2014;96:423–428
159. Haga SB, Burke W, Ginsburg GS, Mills R, Agans R. Primary care physicians' knowledge of and experience with pharmacogenetic testing. Clin Genet 2012;82:388–394
160. Overby CL, Erwin AL, Abul-Husn NS, et al. Physician attitudes toward adopting genome-guided prescribing through clinical decision support. J Pers Med 2014;4:35–49
161. Gerhard GS, Paynton B, Popoff SN. Integrating cadaver exome sequencing into a first-year medical student curriculum. JAMA 2016;315:555–556
162. Manolio TA, Green ED. Leading the way to genomic medicine. Am J Med Genet C Semin Med Genet 2014;166C:1–7
163. Manolio TA, Murray MF; Inter-Society Coordinating Committee for Practitioner Education in Genomics. The growing role of professional societies in educating clinicians in genomics. Genet Med 2014;16:571–572
164. Pearson ER, Donnelly LA, Kimber C, et al. Variation in TCF7L2 influences therapeutic response to sulfonylureas: a GoDARTs study. Diabetes 2007;56:2178–2182
165. Holstein A, Hahn M, Korner A, Stumvoll M, Kovacs P. TCF7L2 and therapeutic response to sulfonylureas in patients with type 2 diabetes. BMC Med Genet 2011;12:30
166. Becker ML, Visser LE, van Schaik RH, Hofman A, Uitterlinden AG, Stricker BH. Genetic variation in the organic cation transporter 1 is associated with metformin response in patients with diabetes mellitus. Pharmacogenomics J 2009;9:242–247
167. Becker ML, Visser LE, van Schaik RH, Hofman A, Uitterlinden AG, Stricker BH. Genetic variation in the multidrug and toxin extrusion 1 transporter protein influences the glucose-lowering effect of metformin in patients with diabetes: a preliminary study. Diabetes 2009;58:745–749
168. Choi JH, Yee SW, Ramirez AH, et al. A common 5′-UTR variant in MATE2-K is associated with poor response to metformin. Clin Pharmacol Ther 2011;90:674–684

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/349303487

Genetic insight into sick sinus syndrome

Article *in* European Heart Journal · February 2021

DOI: 10.1093/eurheartj/ehaa1108

CITATIONS
45

READS
172

44 authors, including:

Rósa Björk Thórólfsdóttir
41 PUBLICATIONS **1,600** CITATIONS
SEE PROFILE

Gardar Sveinbjornsson
deCODE genetics, Inc.
203 PUBLICATIONS **11,078** CITATIONS
SEE PROFILE

Stefania Benonisdottir
University of Oxford
120 PUBLICATIONS **2,937** CITATIONS
SEE PROFILE

Peter E Weeke
Vanderbilt University
179 PUBLICATIONS **10,505** CITATIONS
SEE PROFILE

All content following this page was uploaded by Kaspar R Nielsen on 11 March 2021.

The user has requested enhancement of the downloaded file.

ESC
European Society
of Cardiology

European Heart Journal (2021) **00**, 1–13
doi:10.1093/eurheartj/ehaa1108

CLINICAL RESEARCH
Genetics

Genetic insight into sick sinus syndrome

Rosa B. Thorolfsdottir [1†], Gardar Sveinbjornsson [1†], Hildur M. Aegisdottir[1], Stefania Benonisdottir [1], Lilja Stefansdottir[1], Erna V. Ivarsdottir [1], Gisli H. Halldorsson [1], Jon K. Sigurdsson[1], Christian Torp-Pedersen [2], Peter E. Weeke[3], Søren Brunak[4], David Westergaard [4], Ole B. Pedersen [5], Erik Sorensen[6], Kaspar R. Nielsen [7], Kristoffer S. Burgdorf [6], Karina Banasik [4], DBDS Genomic Consortium[‡], Ben Brumpton [8], Wei Zhou [9], Asmundur Oddsson [1], Vinicius Tragante [1], Kristjan E. Hjorleifsson [1,10], Olafur B. Davidsson[1], Sridharan Rajamani[1], Stefan Jonsson[1], Bjarni Torfason[11,12], Atli S. Valgardsson[12], Gudmundur Thorgeirsson[1,11,13], Michael L. Frigge[1], Gudmar Thorleifsson [1], Gudmundur L. Norddahl[1], Anna Helgadottir[1], Solveig Gretarsdottir[1], Patrick Sulem [1], Ingileif Jonsdottir [1,11,14], Cristen J. Willer [9,15,16], Kristian Hveem[17,18,19], Henning Bundgaard[3], Henrik Ullum [6,20], David O. Arnar[1,11,13], Unnur Thorsteinsdottir [1,11], Daniel F. Gudbjartsson [1,21], Hilma Holm [1*], and Kari Stefansson [1,11*]

[1]deCODE genetics/Amgen, Inc., Sturlugata 8, Reykjavik 101, Iceland; [2]Department of Clinical Research and Cardiology, Nordsjaelland Hospital, Dyrehavevej 29, Hillerød 3400, Denmark; [3]Department of Cardiology, Copenhagen University Hospital, Blegdamsvej 9, Copenhagen 2100, Denmark; [4]Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Blegdamsvej 3A, Copenhagen 2200, Denmark; [5]Department of Clinical Immunology, Naestved Hospital, Ringstedgade 77B, Naestved 4700, Denmark; [6]Department of Clinical Immunology, Copenhagen University Hospital, Blegdamsvej 9, Copenhagen 2100, Denmark; [7]Department of Clinical Immunology, Aalborg University Hospital North, Urbansgade 36, Aalborg 9000, Denmark; [8]Department of Thoracic and Occupational Medicine, St. Olavs Hospital, Trondheim University Hospital, Prinsesse Kristinas gate 3, Trondheim 7030, Norway; [9]Department of Computational Medicine and Bioinformatics, University of Michigan, 100 Washtenaw Avenue, Ann Arbor, MI 48109-2218, USA; [10]Department of Computing and Mathematical Sciences, California Institute of Technology, 1200 E California Blvd. MC 305-16, Pasadena, CA 91125, USA; [11]Faculty of Medicine, University of Iceland, Vatnsmyrarvegur 16, Reykjavik 101, Iceland; [12]Department of Cardiothoracic Surgery, Landspitali—The National University Hospital of Iceland, Hringbraut, Reykjavik 101, Iceland; [13]Department of Medicine, Landspitali—The National University Hospital of Iceland, Hringbraut, Reykjavik 101, Iceland; [14]Department of Immunology, Landspitali—The National University Hospital of Iceland, Hringbraut, Reykjavik 101, Iceland; [15]Department of Internal Medicine: Cardiology, University of Michigan, 1500 East Medical Center Drive, Ann Arbor, MI 48109 -5368, USA; [16]Department of Human Genetics, University of Michigan, 4909 Buhl Building, 1241 E. Catherine St., Ann Arbor, MI 48109 -5618, USA; [17]K.G. Jebsen Center for Genetic Epidemiology, Faculty of Medicine and Health Sciences, Norwegian University of Science and Technology, Erling Skjalgssons gt. 1, Trondheim 7491, Norway; [18]Department of Public Health and Nursing, Faculty of Medicine and Health Sciences, Norwegian University of Science and Technology, Postboks 8905, Trondheim 7491, Norway; [19]HUNT Research Centre, Department of Public Health and General Practice, Norwegian University of Science and Technology, Forskningsveien 2, Levanger 7600, Norway; [20]Statens Serum Institut, Artillerivej 5, Copenhagen 2300, Denmark; and [21]School of Engineering and Natural Sciences, University of Iceland, Hjardarhagi 4, Reykjavik 107, Iceland

Received 28 April 2020; revised 24 August 2020; editorial decision 21 December 2020; accepted 5 January 2021

Aims The aim of this study was to use human genetics to investigate the pathogenesis of sick sinus syndrome (SSS) and the role of risk factors in its development.

Methods and results We performed a genome-wide association study of 6469 SSS cases and 1 000 187 controls from deCODE genetics, the Copenhagen Hospital Biobank, UK Biobank, and the HUNT study. Variants at six loci associated with SSS, a reported missense variant in *MYH6*, known atrial fibrillation (AF)/electrocardiogram variants at *PITX2, ZFHX3, TTN/CCDC141*, and

---

* Corresponding authors. Tel: +354 570 1900, Fax: +354 570 1901, Email: hilma.holm@decode.is (H.H.); Tel: +354 570 1900, Fax: +354 570 1901, Email: kstefans@decode.is (K.S.)

[†] These authors contributed equally to this work.

[‡] DBDS Genomic Consortium: Steffen Andersen, Christian Erikstrup, Thomas F. Hansen, Henrik Hjalgrim, Gregor Jemec, Poul Jennum, Mette Nyegaard, Mie T. Bruun, Mikkel Petersen, Thomas Werge, and Per I. Johansson.

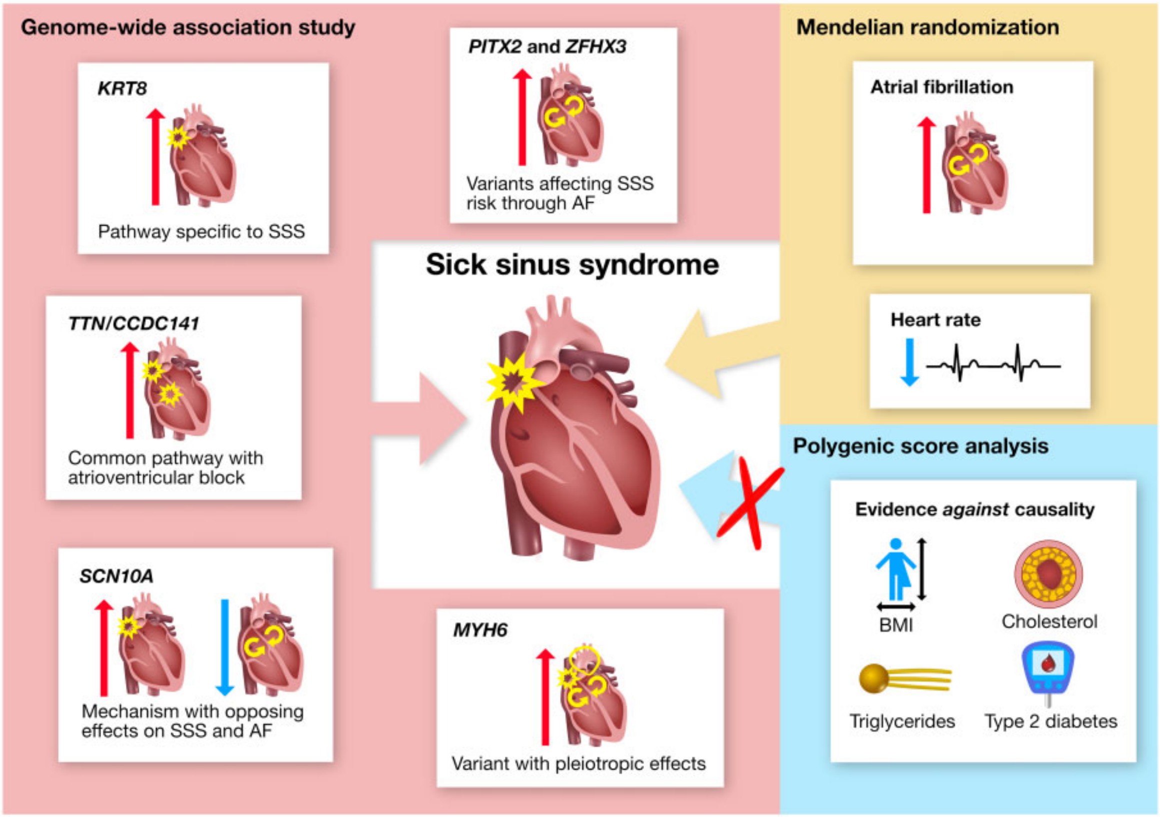
© The Author(s) 2021. Published by Oxford University Press on behalf of the European Society of Cardiology.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from https://academic.oup.com/eurheartj/advance-article/doi/10.1093/eurheartj/ehaa1108/6134552 by Medicinsk Bibliotek, Aalborg Sygehus SYD user on 11 March 2021

SCN10A and a low-frequency (MAF = 1.1–1.8%) missense variant, p.Gly62Cys in *KRT8* encoding the intermediate filament protein keratin 8. A full genotypic model best described the p.Gly62Cys association ($P = 1.6 \times 10^{-20}$), with an odds ratio (OR) of 1.44 for heterozygotes and a disproportionally large OR of 13.99 for homozygotes. All the SSS variants increased the risk of pacemaker implantation. Their association with AF varied and p.Gly62Cys was the only variant not associating with any other arrhythmia or cardiovascular disease. We tested 17 exposure phenotypes in polygenic score (PGS) and Mendelian randomization analyses. Only two associated with the risk of SSS in Mendelian randomization, AF, and lower heart rate, suggesting causality. Powerful PGS analyses provided convincing evidence against causal associations for body mass index, cholesterol, triglycerides, and type 2 diabetes ($P > 0.05$).

Conclusion We report the associations of variants at six loci with SSS, including a missense variant in *KRT8* that confers high risk in homozygotes and points to a mechanism specific to SSS development. Mendelian randomization supports a causal role for AF in the development of SSS.

Graphical Abstract


Genome-wide association study
KRT8
Pathway specific to SSS
PITX2 and ZFHX3
Variants affecting SSS risk through AF
TTN/CCDC141
Common pathway with atrioventricular block
Sick sinus syndrome
SCN10A
Mechanism with opposing effects on SSS and AF
MYH6
Variant with pleiotropic effects
Mendelian randomization
Atrial fibrillation
Heart rate
Polygenic score analysis
Evidence *against* causality
BMI
Cholesterol
Triglycerides
Type 2 diabetes

Summary of genetic insight into the pathogenesis of SSS and the role of risk factors in its development. Variants at six loci (named by corresponding gene names) were identified through GWAS and their unique phenotypic associations provide insight into distinct pathways underlying SSS. Investigation of the role of risk factors in SSS development supported a causal role for AF and heart rate and provided convincing evidence against causality for BMI, cholesterol (HDL and non-HDL), triglycerides and T2D. Mendelian randomization did not support causality for CAD, ischaemic stroke, heart failure, PR interval or QRS duration (not shown in figure). Red and blue arrows represent positive and negative associations, respectively.

Keywords Sick sinus syndrome • GWAS • *KRT8* • Mendelian randomization • Atrial fibrillation

Introduction

Sick sinus syndrome (SSS) is a complex cardiac arrhythmia and the leading indication for permanent pacemaker implantation worldwide.[1] It is characterized by pathological sinus bradycardia, sinoatrial block, or alternating atrial brady- and tachyarrhythmias. Symptoms include fatigue, reduced exercise capacity, and syncope.[2,3] Few studies have been conducted on the basic mechanisms of SSS

Downloaded from https://academic.oup.com/eurheartj/advance-article/doi/10.1093/eurheartj/ehaa1108/6134552 by Medicinsk Bibliotek, Aalborg Sygehus SYD user on 11 March 2021

and therapeutic limitations reflect an incomplete understanding of the pathophysiology. No specific treatment option is aimed at underlying pathways and trying to determine who benefits from cardiac pacing can be challenging.[4] A better understanding of mechanisms leading to SSS is crucial for improvements in treatment and prevention and genetic studies provide an opportunity to gain such insight.

Coding variants in several genes, including *HCN4, SCN5A*, and *GNB2*, have been implicated in rare familial SSS through linkage analysis and candidate gene methods.[5–15] In the only published genome-wide association study (GWAS) of SSS to date, we identified a strong association with a missense variant in *MYH6*, encoding the alpha-heavy chain subunit of cardiac myosin. This was the first GWAS discovery implicating a cardiac structural or contractile unit as a primary cause of arrhythmia.[16]

Age is the strongest risk factor for SSS,[17] potentially reflecting degenerative fibrosis and electrical remodelling of the SA node and the atria in general.[18–20] SSS often coexists with the more common atrial fibrillation (AF)[21] and the two arrhythmias are thought to predispose to each other.[22] We have previously examined the role of AF in SSS development through Mendelian randomization,[23] a method using sequence variants associated with a risk factor (AF) as unbiased proxy indicators to determine whether the risk factor can cause a disease (SSS).[24,25] The study supported causality,[23] likely mediated through atrial remodelling.[26,27] Several other traits have been associated with SSS while the nature of the associations has not been ascertained. These include body mass index (BMI), height, heart rate, and cardiovascular diseases such as hypertension, myocardial infarction, heart failure, and stroke.[17]

To gain insight into the pathogenesis of SSS, we performed a large GWAS of over 6000 cases and one million controls of European descent. We used polygenic scores (PGSs) and Mendelian randomization to determine the nature of the associations of risk factors with SSS and in particular whether they contribute to the cause of the disease.

Methods

The study design is two-fold (Supplementary material online, *Figure S1*). First, we performed a meta-analysis of GWASs, searching for associations of sequence variants with SSS. In total 6469 SSS cases were compared to 1 000 187 controls in a meta-analysis of material from deCODE genetics Iceland, the Copenhagen Hospital Biobank Cardiovascular Study (CHB-CVS)/Danish Blood Donor Study (DBDS), and the UK Biobank,[28] with follow-up in the Norwegian Nord-Trøndelag Health Study (HUNT).[29] Second, we performed PGS analysis and Mendelian randomization to examine the role of putative risk factors in SSS development (for detailed methods, see Supplementary material online).

Sick sinus syndrome study populations

The deCODE genetics SSS sample consisted of 3577 Icelanders diagnosed with SSS and 347 764 controls. The CHB-CVS included 2209 SSS cases. The control group included blood donors from the DBDS ($N > 99\ 000$)[30] and SSS cases were also compared to subjects in CHB-CVS with other cardiovascular conditions ($N > 89\ 000$). The SSS population from the UK Biobank consisted of 403 cases and 403 181 controls.[28] The HUNT cohort consisted of 280 Norwegian SSS cases and 69 141 controls recruited through a population-based health survey conducted in the Nord-Trøndelag County, Norway.[29] In all four cohorts, SSS diagnosis was based on International Classification of Diseases 9th (ICD-9: 427.8) or 10th revision (ICD-10: I49.5) from hospitals and/or outpatient clinics. The cohorts are described in more detail in Supplementary material online, Methods and Supplementary material online, *Table S1*.

Genotyping

The deCODE study was based on whole-genome sequence data from 28 075 Icelanders participating in various disease projects. The 32.9 million variants that passed quality threshold were imputed into 127 175 Icelanders who had been genotyped using Illumina SNP chips. Finally, genotype probabilities for untyped relatives were calculated based on Icelandic genealogy.[31] Genotyping for CHB-CVS and DBDS was performed at deCODE using a north European sequencing panel of 15 576 individuals (including 8429 Danes) for imputation into those chip typed (methods manuscript in preparation). Genotyping in the UK Biobank has been described in detail elsewhere.[32–35] Illumina HumanCoreExome arrays were used for genotyping in the HUNT cohort.

Genome-wide association study: statistical analysis

We performed a meta-analysis of GWAS of 6189 SSS cases and 931 046 controls from deCODE genetics, CHB-CVS/DBDS, and the UK Biobank. We then tested genome-wide significant and suggestive variants (Supplementary material online, Methods) among 280 cases and 69 141 controls from the Norwegian HUNT study. We used logistic regression to test for association between sequence variants and SSS and other case–control phenotypes, treating phenotype status as the response and allele count as a covariate. Other available individual characteristics that correlate with phenotype status were included in the model as nuisance variables. SSS associations were tested under additive, recessive, and full genotypic models. The full genotypic model includes separate parameters for heterozygotes and homozygotes. We corrected the threshold for genome-wide significance for multiple testing with a weighted Bonferroni adjustment using as weights the enrichment of variant classes with predicted functional impact among association signals estimated from the Icelandic data.[36] Significance thresholds were 1.3×10^{-7} for variants with high impact, 2.6×10^{-8} for variants with moderate impact (including missense), 2.4×10^{-9} for low-impact variants, 1.2×10^{-9} for other variants in Dnase I hypersensitivity sites, and 7.5×10^{-10} for all other variants.[36] We used linear regression to test variant associations with quantitative phenotypes, treating the quantitative measurement as response and the genotype as covariate. These included endophenotypes of SSS, chronotropic response to exercise ($N = 7746$), and electrocardiogram (ECG) measurements ($N \sim 73\ 000$) (see Supplementary material online, Methods for detail).

Analysis of genetic risk shared by sick sinus syndrome and putative risk factors

We used PGSs for 17 exposure phenotypes to examine their correlation with SSS (Supplementary material online, *Table S2*). The phenotypes were chosen because of reported epidemiological associations with SSS or other cardiovascular conditions (see Supplementary material online, Methods). The PGSs were generated using summary statistics from the largest available GWASs (training sample) for each phenotype that do not include deCODE data (Supplementary material online, *Table S2* and Supplementary material online, *Figure S1*). Subsequently, they were tested for association with the risk of SSS among 2556 chip-typed individuals from deCODE (target sample). *P*-values were corrected using genomic control.[37]

The use of PGSs to detect association between exposure and outcome is a robust method such that the absence of association in a well-powered PGS analysis provides strong evidence against causality.[25]

Downloaded from https://academic.oup.com/eurheartj/advance-article/doi/10.1093/eurheartj/ehaa1108/6134552 by Medicinsk Bibliotek, Aalborg Sygehus SYD user on 11 March 2021

Table 1 Association results for lead variants at loci reaching genome-wide significance in a meta-analysis of sick sinus syndrome including 6469 cases and 1 000 187 controls

Locus number	Rs-name/Chr: position (hg38)	Effect allele/other	EAF^a (%)	Variant annotation	Coding change	Closest gene	OR (95% CI)	*P*-value	*P*-value threshold
1	rs387906656^b/chr14:23396970	A/G	0.34	Missense	p.Arg721Trp	*MYH6*	8.88 (6.97–11.32)	7.5×10^{-70}	2.6×10^{-8}
2	rs7689774/chr4:110782354	T/G	20.07	Intergenic	–	*MIR297, PITX2*	1.21 (1.15–1.26)	2.0×10^{-15}	1.2×10^{-9}
3	rs11554495/chr12:52904798	A/C	1.64	Missense	p.Gly62Cys	*KRT8*	1.62 (1.43–1.84)	9.4×10^{-14}	2.6×10^{-8}
4	rs12932445/chr16:73035989	C/T	20.10	Intronic	–	*ZFHX3*	1.16 (1.11–1.21)	8.1×10^{-10}	1.2×10^{-9}
5	rs35813871^c/chr2:178785681	A/G	26.63	Missense	p.Thr811Ile	*TTN*	1.13 (1.09–1.18)	5.7×10^{-9}	2.6×10^{-8}
	rs34883828^c/chr2:178905448	A/C	15.23	Missense	p.Glu382Asp	*CCDC141*	1.15 (1.09–1.21)	1.1×10^{-7}	2.6×10^{-8}
6	rs6795970/chr3:38725184	A/G	35.78	Missense	p.Val1073Ala	*SCN10A*	1.12 (1.07–1.16)	2.5×10^{-8}	2.6×10^{-8}

CI, confidence interval; OR, odds ratio.

[a] Effect allele frequency in Iceland (deCODE).

[b] Variant exclusive to Iceland.

[c] p.Thr811Ile in *TTN* and p.Glu382Asp in *CCDC141* are weakly correlated: $R^2 = 0.20$, $D' = 0.64$.

However, finding association in PGS analysis does not confirm causality.[38,39] For the eight PGSs that associated with SSS, we performed a 2-sample Mendelian randomization analysis, equivalent to a fixed effect MR-Egger[40] using published genome-wide significant SNPs from the largest available GWAS on each exposure phenotype as genetic instruments.[41–49] To estimate the causal effect, we regressed the published effects of the SNPs on the respective exposure phenotype against their SSS effects in deCODE, CHB-CVS/DBDS, and the UK Biobank, using minor allele frequency (MAF) × (1 - MAF) as weights. For the ECG measurements PR interval and QRS duration, effects in milliseconds were converted to standard deviations (SDs) according to one SD in the Icelandic data so that all causal estimates in a Mendelian randomization forest plot correspond to equal odds ratio (OR) or SD changes for binary and quantitative traits, respectively. For significant associations, we performed further sensitivity analysis to detect outliers (funnel plots) and tested for directional pleiotropy using the MR-Egger intercept test (see Supplementary Material online, Methods).[40]

Results

Six sick sinus syndrome loci

In a GWAS comparing 6469 SSS cases to 1 000 187 controls, variants at six loci satisfied our criteria for genome-wide significance (*Table 1*, Supplementary material online, *Table S3*, Figure 1, and Supplementary material online, *Figure S2*), with ORs ranging from 1.12 to 8.88. One is at a novel SSS locus, a low-frequency missense variant, p.Gly62Cys (MAF = 1.1–1.8% in the four cohorts) in the gene *KRT8* on chromosome 12. This variant has not been associated with arrhythmias or ECG traits before. The other variants are the rare Icelandic SSS missense variant p.Arg721Trp in the sarcomere gene *MYH6*,[16] common variants at two AF loci, *PITX2*[50] and *ZFHX3*,[51] and two ECG loci, *TTN/CCDC141*[52] and *SCN10A*.[52–54] We have previously reported secondary associations of variants at the four AF and ECG loci with SSS.[23,52] At the *TTN/CCDC141* locus, we report associations of two missense variants, p.Thr811Ile in *TTN* and p.Glu382Asp in *CCDC141*, that are weakly correlated ($R^2 = 0.20$, $D' = 0.64$, Supplementary material online, *Table S4*).

High penetrance of sick sinus syndrome among homozygotes for p.Gly62Cys in *KRT8*

The SSS association of p.Gly62Cys in *KRT8* was discovered with the additive model (*Table 1* and Supplementary material online, *Tables S5 and S6*). However, homozygotes of p.Gly62Cys have a higher risk of SSS than assumed under the additive model (*Table 2*, homozygous genotypic OR > $1.62^2 = 2.62$, $P_{\text{full vs. additive}} = 2.1 \times 10^{-8}$). We observed an OR of 1.44 for heterozygotes (95% CI = 1.26–1.65) and 13.99 for homozygotes (95% CI = 8.16–23.98), compared to non-carriers (*P*-value for the full genotypic model = 1.6×10^{-20}, *Table 2*). None of the other SSS associations deviated from the additive model ($P > 0.05$). *KRT8* encodes the intermediate filament keratin 8 (K8, Supplementary material online, *Figure S3*) which is widely expressed, including in the heart,[55] and we detected its expression (52.1 ± 30.2 transcripts per million) in our cardiac samples from right atria ($n = 169$, Supplementary material online, Methods). Neither p.Gly62Cys nor two correlated variants ($R^2 > 0.6$) associated with

Downloaded from https://academic.oup.com/eurheartj/advance-article/doi/10.1093/eurheartj/ehaa1108/6134552 by Medicinsk Bibliotek, Aalborg Sygehus SYD user on 11 March 2021

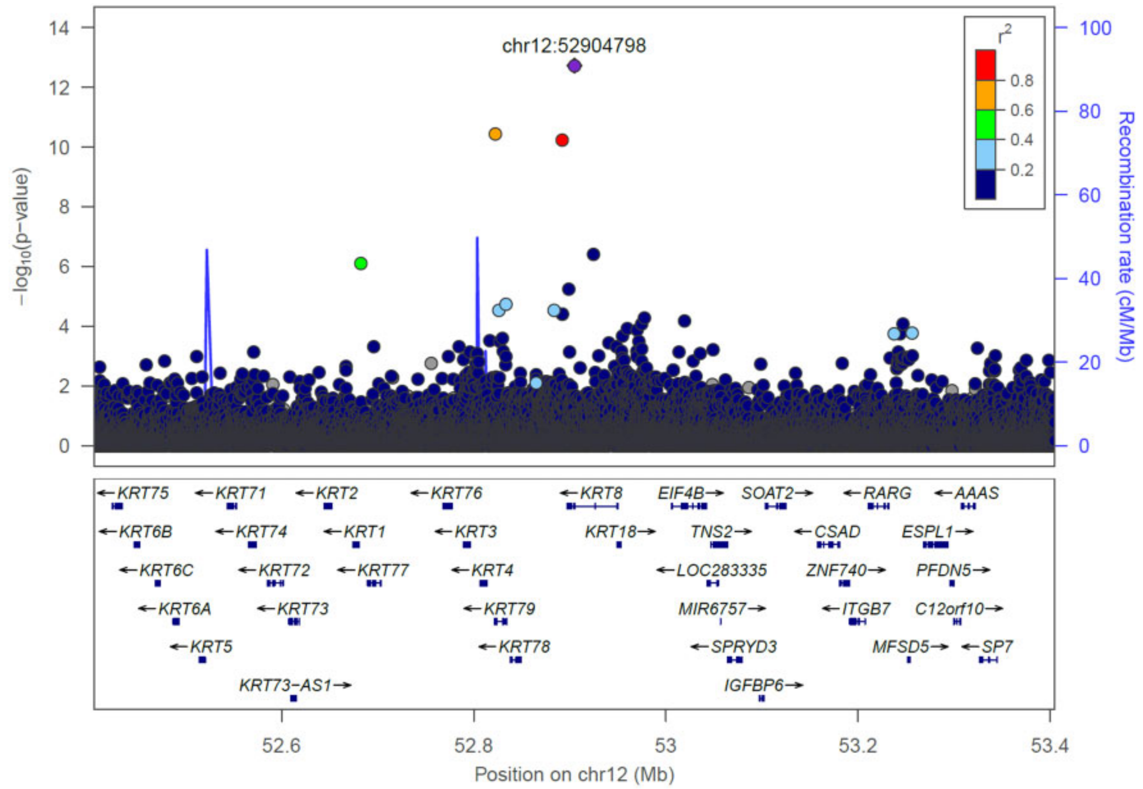


Figure 1 Regional plot of the *KRT8* locus on chromosome 12q13. The plot depicts association results (*P*-values) with SSS ($N = 6189$) in a meta-analysis with data from deCODE, the CHB-CVS/DBDS, and the UK Biobank. The *y*-axis shows the -$\log_{10}$ *P*-value and *x*-axis shows the genomic position (hg38). The lead variant of the signal (p.Gly62Cys) is labelled as a diamond and coloured purple. Other variants are coloured according to correlation (R^2) with the lead variant in the deCODE data. The plot includes variants common to the three datasets as well as variants specific to the Icelandic deCODE data.

the expression of *KRT8* or nearby genes (1 Mb) in our cardiac samples (Supplementary material online, *Table S7*) or in GTEx.[55]

About 1 in 5000 individuals from the four European populations in this study is homozygous for p.Gly62Cys in *KRT8*, consistent with the Hardy–Weinberg equilibrium ($P = 0.071$–0.82). To evaluate the penetrance of SSS among homozygous carriers of p.Gly62Cys, we analysed the two largest SSS datasets in the study, deCODE and CHB-CVS/DBDS. Considering that SSS is a late onset disease, the observed penetrance is relatively high although incomplete (*Figure 2*). Eight out of 36 genotyped Icelandic homozygotes (aged 23–88 years, mean 59.9) had SSS (Supplementary material online, *Table S8*) and 10 out of 79 homozygotes (aged 27–103 years, mean 62.1) in CHB-CVS/DBDS. Homozygotes were not significantly younger at SSS diagnosis than heterozygotes or non-carriers ($P = 0.53$ in deCODE and $P = 0.26$ in CHB-CVS/DBDS, Supplementary material online, *Table S9*).

Diverse associations of sick sinus syndrome variants with other phenotypes

We tested the associations of the SSS variants with other phenotypes in deCODE, CHB-CVS/DBDS, and the UK Biobank datasets (Supplementary material online, *Tables S10–S14*). In addition to SSS, p.Gly62Cys in *KRT8* associated with pacemaker implantation (OR heterozygotes = 1.28, 95% CI = 1.13–1.45, OR homozygotes = 9.17, 95% CI = 2.89–29.09, $P = 1.9 \times 10^{-8}$). The variant did not associate with other arrhythmias or cardiovascular diseases, with electrolyte or hormonal disturbances that are linked to the development of SSS (potassium, calcium, thyroid stimulating hormone)[57] or with the suggested consequence of p.Gly62Cys in candidate gene studies[58] (pancreatitis/lipase, liver disease, Bonferroni-adjusted significance threshold of $P = 0.05/23 = 0.0022$, Supplementary material online, *Tables S10 and S11*). Furthermore, p.Gly62Cys did not associate with any of the 122 ECG parameters (*N* up to 72 825, Supplementary material online, *Table S13*).

The associations of SSS variants with AF varied. p.Gly62Cys in *KRT8* (Supplementary material online, *Tables S10 and S11*) and the missense variants in *TTN*/*CCDC141* (Supplementary material online, *Table S12*) did not associate with the risk of AF, applying a Bonferroni-corrected significance threshold of $P < 0.05/14 = 0.0036$. The *TTN*/*CCDC141* variants are located within 300 kbp from previously reported AF variants[41] and are independent of them ($R^2 < 0.2$, Supplementary material online, *Figure S4*). The other four SSS loci have been reported to affect AF but their pattern of association with the two phenotypes differ. p.Arg721Trp in *MYH6*[23] associates stronger with SSS than AF but the reverse applies to variants at *PITX2*. At

Downloaded from https://academic.oup.com/eurheartj/advance-article/doi/10.1093/eurheartj/ehaa1108/6134552 by Medicinsk Bibliotek, Aalborg Sygehus SYD user on 11 March 2021

Table 2 **Association of p.Gly62Cys in *KRT8* with sick sinus syndrome in the deCODE, CHB-CVS/DBDS, UK Biobank, and HUNT samples under the full genotypic model, calculating independent risk among heterozygotes and homozygotes**

Rs-name/Pos (hg38)	Effect allele/other	Coding change	Gene	Cohort	*N*	EAF (%)	Risk for heterozygous carriers, OR (95% CI)	Risk for homozygous carriers, OR (95% CI)	*P*-value for full genotypic model[a]
rs11554495/chr12:52904798	A/C	p.Gly62Cys	*KRT8*	deCODE	3577	1.64	1.42 (1.17–1.72)	12.95 (4.58–36.63)	1.5×10^{-9}
				CHB-CVS	2209	1.77	1.34 (1.08–1.66)	16.13 (6.24–41.67)	2.6×10^{-10}
				UK Biobank	403	1.05	2.35 (1.42–3.88)	17.90 (0.74–435.16)	0.00012
				HUNT	280	1.15	1.55 (0.77–2.91)	0.00[b]	0.50
				Combined	6469	–	1.44 (1.26–1.65)	13.99 (8.16–23.98)	1.6×10^{-20}

CI, confidence interval; EAF, effect allele frequency; OR, odds ratio; Pos, position; SSS, sick sinus syndrome.

[a] The full genotypic model ($P_{\text{full}} = 1.6 \times 10^{-20}$) deviates significantly from the additive model ($P_{\text{additive}} = 9.4 \times 10^{-14}$). *P* for deviation from the additive model = 2.1×10^{-8}.

[b] Among 14 homozygotes in the HUNT study, none had SSS (*P*-value for homozygotes = 0.97).

ZFHX3, the effects on AF and SSS are comparable.[41,50,51,59,60] p.Val1073Ala in *SCN10A* is one of the top AF variants at this locus (Supplementary material online, *Figure S5*), but the allele that increases the risk of SSS protects against AF (Supplementary material online, *Table S12*).[53,61]

All the SSS variants increased the risk of pacemaker implantation (Supplementary material online, *Tables S10–S12*), as expected since SSS is the most common reason for this procedure. We have previously described in detail the associations of p.Arg721Trp in *MYH6* with coarctation of the aorta, other congenital malformations of the heart, aortic valve stenosis, and heart failure[23,62,63] (Supplementary material online, *Table S12*). We note that p.Glu382Asp in *CCDC141* associated with both second- and third-degree atrioventricular block (AVB)[52] and the association with third-degree AVB was genome-wide significant ($P = 1.3 \times 10^{-14}$, OR = 1.27, 95% CI = 1.20–1.35) and more significant than the SSS association. Another notable association is that of the *PITX2* variant with heart failure (Supplementary material online, *Table S12*) as recently reported in a GWAS of heart failure.[47] None of the SSS variants associated with chronotropic response to exercise (Supplementary material online, *Tables S15 and S16*) in our dataset. However, the *SCN10A* and *CCDC141* loci have been reported to associate with heart rate profile during exercise in larger studies of UK Biobank data.[64,65]

Genetically predicted atrial fibrillation and lower heart rate associate with sick sinus syndrome

PGSs are powerful tools for detecting associations between phenotypes.[38] We generated PGSs for 17 exposure phenotypes using summary statistics from the largest available GWASs (Supplementary material online, *Table S2*) and tested them for association with SSS in deCODE data. Eight PGSs, for AF, heart rate, coronary artery disease (CAD), height, QRS duration, PR interval, ischaemic stroke, and heart failure, associated with SSS while those for BMI, type 2 diabetes (T2D), non-HDL cholesterol, HDL cholesterol, triglycerides, and hypertension/blood pressure traits did not (Bonferroni-adjusted significance threshold of $P = 0.05/17 = 0.0029$, Supplementary material online, *Table S17*). The PGS for heart rate was the only one to inversely associate with SSS risk.

We then performed Mendelian randomization to determine the causality of the associations between the eight exposures and SSS, using genome-wide significant SNPs from GWAS of the exposure phenotypes as instruments (see Methods). Genetic predisposition to AF and genetically determined heart rate and height associated with the risk of SSS but the height association did not remain after accounting for AF in a multivariate analysis (*Figures 3* and *4A* and *B*).[66–68] With few exceptions, the effects of AF variants on SSS are proportional to their effects on AF ($P = 7.8 \times 10^{-14}$, *Figure 4A* and Supplementary material online, *Figure S6A*). The greatest deviation from the expected SSS effect was for the association at *SCN10A* (tagged by rs6790693, OR for AF = 1.06, OR for SSS = 0.89, Supplementary material online, *Figure S6A*). Mendelian randomization did not support causality for CAD, ischaemic stroke, heart failure, PR interval, or QRS duration (*Figure 3* and Supplementary material online, *Figure S7*).

Downloaded from https://academic.oup.com/eurheartj/advance-article/doi/10.1093/eurheartj/ehaa1108/6134552 by Medicinsk Bibliotek, Aalborg Sygehus SYD user on 11 March 2021

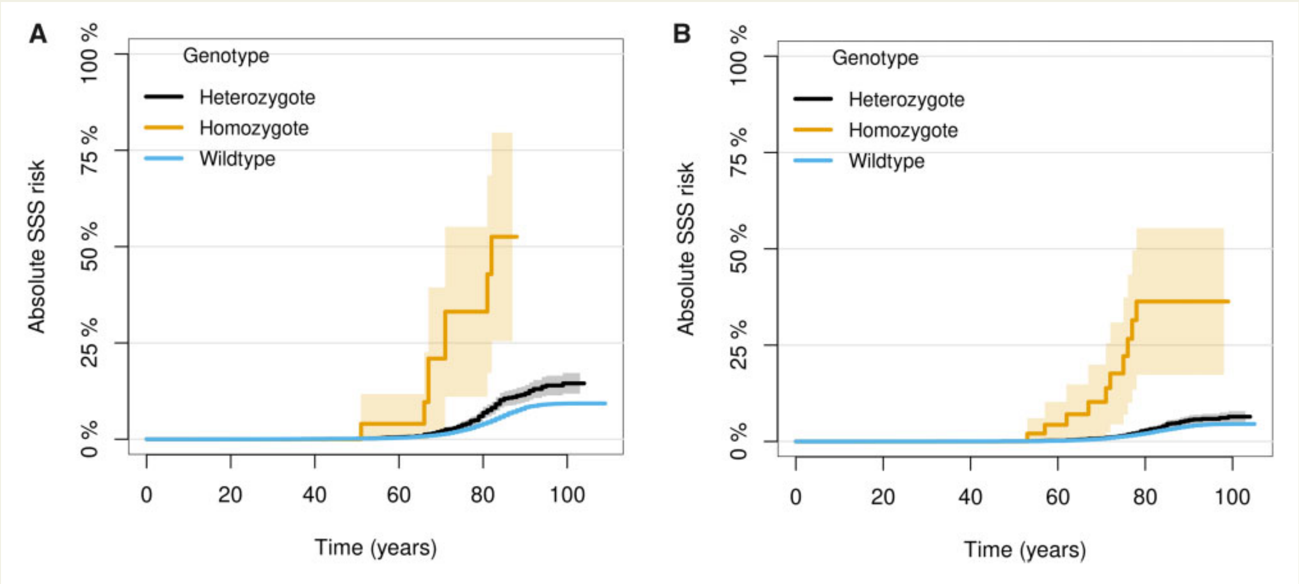


Figure 2 Cumulative incidence curves for SSS age of onset based on *KRT8* p.Gly62Cys genotype (*A*) in deCODE and (*B*) in CHB-CVS/DBDS. Aalen–Johansen estimator was used, treating death as a competing risk.[56] Age of onset was determined by the first registered ICD diagnosis of SSS available to us and thus represents an upper range (Supplementary material online, Methods).

Discussion

Traditionally, SSS has been considered as a collection of conditions with a variety of causes, intrinsic and extrinsic to the SA node.[69] In line with this notion, we identified sequence variants associating with SSS at six loci implicating several distinct pathways in SSS development. One of the SSS associations is novel, with a low-frequency missense variant, p.Gly62Cys in *KRT8*. One is with the reported p.Arg721Trp in *MYH6*,[16] and four confirm loci previously implicated in SSS secondarily to AF, at *PITX2*,[50] and *ZFHX3*,[51] or ECG traits, at *SCN10A*[53] and *TTN/CCDC141*.[52] The leading variants at four of the SSS loci are missense.

KRT8 encodes K8, a cytoskeletal intermediate filament protein widely expressed, including in skeletal muscle and the heart.[55] Historically, keratins have been described as epithelial-specific intermediate filaments[70] and mutations affecting keratin function have almost exclusively been associated with diseases of the skin and other epithelial tissues.[71,72] The association of p.Gly62Cys in *KRT8* with SSS described here supports the notion of a role for K8 in the human heart, as previously suggested by animal studies.[73–76] Experiments have revealed loss-of-function effects of p.Gly62Cys, that is located in the head domain of K8 (Supplementary material online, *Figure S3*).[77,78] Glycine at position 62 is a conserved amino acid (GERP[79] score 1.73) and the variant introduces cysteine into a protein otherwise devoid of cysteine residues. This interrupts intermediate filament assembly *in vitro*[77] and prevents K8 from serving as a phosphate 'sponge' for stress-activated kinases that mediate apoptosis.[78] Thus, the variant could increase the risk of SSS either by affecting the structural or cardioprotective role of K8.[77,78]

In addition to p.Gly62Cys in *KRT8*, lead variants at two other SSS loci are missense variants in genes encoding structural components of the heart. *MYH6* encodes the alpha-myosin heavy chain (αMHC) and *TTN* encodes the giant protein titin, both of which are key components of the sarcomere.[80,81] Since the identification of *MYH6* in

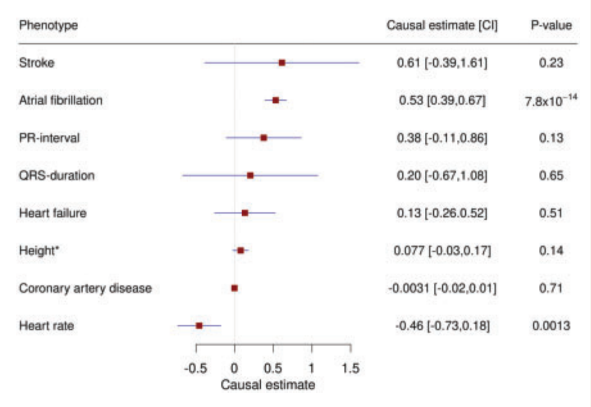


Figure 3 Forest plot showing causal effect estimates from a Mendelian randomization analysis of putative risk factors on SSS. We regressed published effects of SNPs on the respective exposure phenotype against their SSS effects in deCODE, CHB-CVS/DBDS, and the UK Biobank. The causal estimates are the slopes from the regression lines where a one unit change in the exposure phenotypes equals a one standard deviation change for quantitative traits and an odds ratio of 2.7 for binary exposure phenotypes. The forest plot shows the causal estimates, 95% confidence intervals, and corresponding *P*-values from the Mendelian randomization analysis. *For height, the result from a multivariate analysis accounting for AF effects is shown.

Downloaded from https://academic.oup.com/eurheartj/advance-article/doi/10.1093/eurheartj/ehaa1108/6134552 by Medicinsk Bibliotek, Aalborg Sygehus SYD user on 11 March 2021

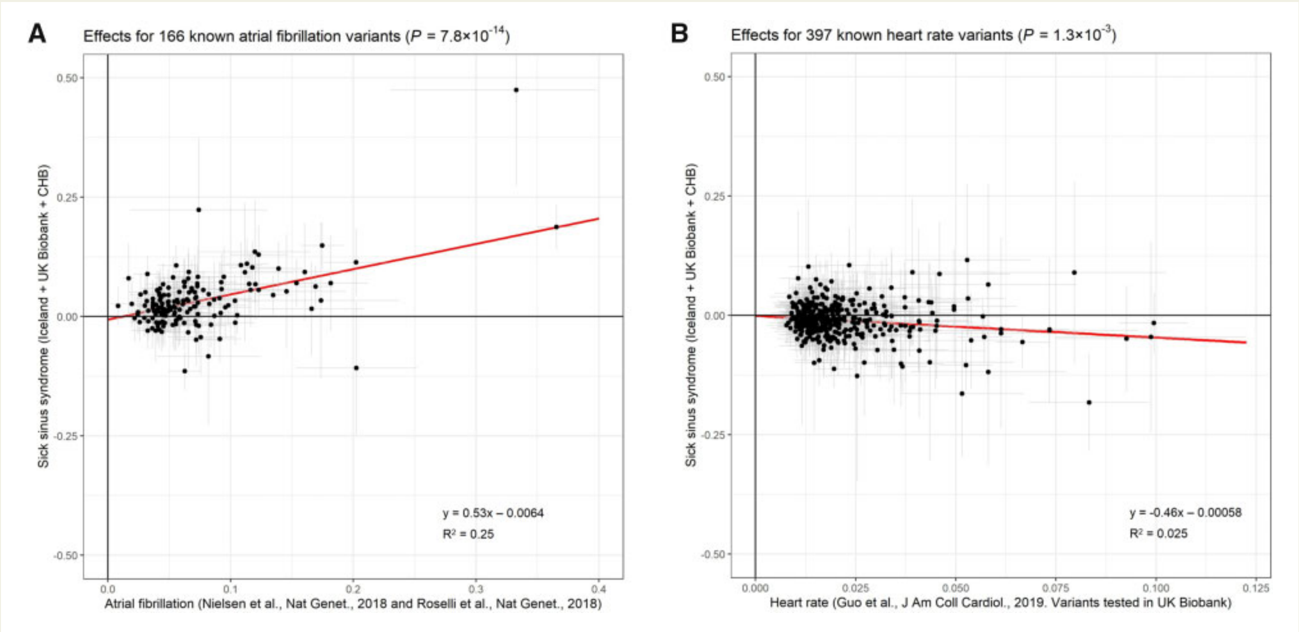


Figure 4 Visualization of significant associations from Mendelian randomization analysis. Effects of published variants on two phenotypes plotted against their effects on SSS in deCODE, CHB-CVS/DBDS, and the UK Biobank (*y*-axes, $N = 6189$). The *x*-axes depict (*A*) published AF effects[41,42] and (*B*) heart rate effects in UK Biobank.[43] The equations for the regression lines are shown and the coefficients of determination (R^2).

the first GWAS of SSS,[16] structural components of cardiomyocytes have increasingly been implicated in atrioventricular conduction and arrhythmias, in particular AF.[23,41,82,83] However, the variants at *KRT8, MYH6* and *TTN* do not mediate their risk through AF because they either do not associate with AF (*KRT8* and *TTN*) or have a stronger effect on SSS risk (*MYH6*).

K8, αMHC, and titin have a common role in cardiac adaptive responses to stress which may be of relevance to their association with SSS. αMHC allows greater economy in force generation than the homologous beta chain and a shift in their relative expression plays a major role in regulating myocardial contractile activity.[84,85] Likewise, the isoform switch of titin alters ventricular filling[81,86] and K8 is upregulated under stress to maintain structural integrity.[76] The expression of all three genes is modified in human heart failure.[76,84,87] Thus, whether or not the SSS variants in *KRT8, MYH6* and *TTN* directly affect the SA node, it is possible that they could contribute to inadequate cardiac output and symptom development in SSS by interrupting adaptive increase in stroke volume.

The SSS variants at *PITX2* and *ZFHX3* point to genes encoding transcription factors.[88,89] They likely mediate their effects on SSS through AF since these have consistently been the strongest AF loci in GWAS.[42,50,51] Conversely, p.Val1073Ala in *SCN10A* has opposite effects on the two arrhythmias. The allele that increases the risk of SSS protects against AF.[53,61] It also prolongs the PR interval[53,54,61] and duration of the QRS complex.[52,54] *SCN10A* encodes the neuronal sodium channel isoform $\mathrm{Na_v}$1.8. A plausible mechanism behind the unique phenotypic effects of p.Val1073Ala involves $\mathrm{Na_v}$1.8s role in the way in which the autonomic nervous system affects the heart.[90–92] In support of this, animal studies have shown that blockade of $\mathrm{Na_v}$1.8 in intracardiac autonomic neurons in the ganglionated plexi (GP) suppresses AF inducibility.[90,91] The same applies to GP ablation in patients with AF.[93–95] However, GP ablation is controversial due to observed adverse events,[96–98] including development of SSS and pacemaker implantation.[99] The opposite effects of p.Val1073Ala on SSS and AF risk support concerns that GP ablation for AF treatment may cause SSS and links this adverse effect to $\mathrm{Na_v}$1.8 function specifically.

Our genetic analyses shed light on pathways contributing to the development of SSS and the role of risk factors. In particular, they provide insight into the intricate and debated relationship between SSS and AF.[22] Our Mendelian randomization analysis reveals a strong association between genetic predisposition for AF and risk of SSS. This is consistent with our previous report[23] and likely reflects causality. Pacing-induced AF has been shown to impair SA node function in dogs[100] and AF causes regional atrial substrate changes around the SA node in humans.[27] Pathobiological pathways common to AF and SSS, such as inflammation, interstitial atrial fibrosis, and alterations in intracellular Ca^{2+} dynamics, may also contribute to this trend.[22,26,101]

Importantly, our results also show that SSS does not only occur as a consequence of AF or pathways common to AF. This is evident in the case of SSS variants at *KRT8* and *TTN/CCDC141* that do not associate with the risk of AF in a large dataset. p.Gly62Cys in *KRT8* is the only SSS variant that does not associate with other arrhythmias, ECG traits, or cardiovascular diseases in our dataset. Thus, this variant in particular may point to a mechanism that is specific to SSS development. We observe a genome-wide significant association of p.Glu382Asp in *CCDC141* with third-degree AVB, the first one reported for complete heart block. Therefore, this locus points to a

Downloaded from https://academic.oup.com/eurheartj/advance-article/doi/10.1093/eurheartj/ehaa1108/6134552 by Medicinsk Bibliotek, Aalborg Sygehus SYD user on 11 March 2021

mechanism linking the functions of the AV and SA nodes and does not mediate the risk of SSS as a consequence of AF. The complexity of the relationship between SSS and AF is further evident by the opposite effects of the *SCN10A* locus on the two arrhythmias. Finally, since SSS variants do not consistently associate with AF, our data do not suggest a strong causal role for SSS in the development of AF.

Heart rate variants associate inversely with SSS, consistent with epidemiological observation,[17] potentially because bradycardia is an integral part of the SSS diagnosis. However, lower heart rate could also represent a direct marker of biological processes affecting SSS, such as fibrosis or altered autonomic neural input. Finally, lack of associations with powerful PGSs based on large datasets serves as convincing evidence against causality.[25] In particular, this applies to BMI, cholesterol, triglycerides, and T2D.

Several of our findings have direct clinical implications. In particular, the causal role of AF in SSS development emphasizes the need for heightened awareness of potential SSS among patients with AF, especially those with unspecific symptoms. Furthermore, the strong association of the SSS variant p.Glu382Asp in *CCDC141* with AVB encourages consideration of dual chamber pacemaker implantation in SSS patients carrying this variant. Lastly, the *SCN10A* locus points to SSS as a potential adverse effect of GP ablation therapy for AF.

The major strength of the study is the large SSS sample set and extensive phenotypic and genotypic data for the same datasets. Different proportions of comorbidities and drug use in SSS cases and controls could constitute confounding; however, the consistent associations of SSS variants with both SSS and pacemaker implantation across cohorts support their true effect on cardiac conduction. Furthermore, if a specific comorbidity would explain one of the SSS associations, this would be evident by a stronger effect on that phenotype. The strength of Mendelian randomization in determining causality includes less concern for confounding and reverse causation than in observational studies since sequence variants associated with the exposure (e.g. AF) are randomized during meiosis and this process is unaffected by the presence of the outcome (SSS) later in life.[24] For heart failure, a limitation of the Mendelian randomization analysis is that variants identified in GWAS are not ideal instruments because of their association with potential confounders (e.g. AF, CAD, and BMI).[47]

Conclusion

In this large genetic study, we found six SSS loci, some of which also associate with other arrhythmias and cardiac electrical function. One of the associations is with a missense variant in the intermediate filament gene *KRT8* that confers high risk of SSS among homozygous carriers. Mendelian randomization analysis suggests that AF and lower heart rate are directly involved in SSS development. PGS analysis provides convincing evidence against causality for BMI, cholesterol, triglycerides, and T2D. On the other hand, the data also show that SSS can result from perturbation of pathways unrelated to AF. In particular, p.Gly62Cys at *KRT8* does not associate with any other cardiovascular traits and points to a mechanism that is specific to SSS development.

Supplementary material

Supplementary material is available at *European Heart Journal* online.

Data availability

The Icelandic population WGS data has been deposited at the European Variant Archive under accession code PRJEB15197. We declare that the data supporting the findings of this study are available within the article, its Supplementary material online and upon reasonable request. The genome-wide association scan summary data will be made available at http://www.decode.com/summarydata.

Acknowledgements

We thank all the study subjects for their valuable participation as well as our colleagues that contributed to data collection, sample handling, and genotyping.

Funding

This work was partly supported by NordForsk through the funding to PM Heart, project number 90580, the Innovation Fund Denmark (IFD) under File No. 8114-00033B and the Technology Development Fund, Iceland, project number 90580.

Conflict of interest: The following authors affiliated with deCODE genetics/Amgen, Inc., are employed by the company: R.B.T., G.S., H.M.A., S.B., L.S., E.V.I., G.H.H., J.K.S., A.O., V.T., K.E.H., O.B.D., S.R., S.J., G.T., M.L.F., G.T., G.L.N., A.H., S.G., P.S., I.J., D.O.A., U.T., D.F.G., H.H., and K.S.. D.W. received grants from Novo Nordisk Foundation during the conduct of the study. S.B. is a board member for Proscion A/S and Intomics A/S.

References

1. Mond HG, Proclemer A. The 11th world survey of cardiac pacing and implantable cardioverter-defibrillators: calendar year 2009—a World Society of Arrhythmia's project. *Pacing Clin Electrophysiol* 2011;**34**:1013–1027.
2. Ferrer MI. The sick sinus syndrome. *Circulation* 1973;**47**:635–641.
3. Fuster V, Harrington RA, Narula J, Eapen ZJ, *Hurst's the Heart.* 14th ed. New York: McGraw-Hill Education; 2017.
4. Brignole M, Auricchio A, Baron-Esquivias G, Bordachar P, Boriani G, Breithardt OA, Cleland J, Deharo JC, Delgado V, Elliott PM, Gorenek B, Israel CW, Leclercq C, Linde C, Mont L, Padeletti L, Sutton R, Vardas PE. 2013 ESC guidelines on cardiac pacing and cardiac resynchronization therapy: the task force on cardiac pacing and resynchronization therapy of the European Society of Cardiology (ESC). Developed in collaboration with the European Heart Rhythm Association (EHRA). *Europace* 2013;**15**:1070–1118.
5. Le Scouarnec S, Bhasin N, Vieyres C, Hund TJ, Cunha SR, Koval O, Marionneau C, Chen B, Wu Y, Demolombe S, Song LS, Le Marec H, Probst V, Schott JJ, Anderson ME, Mohler PJ. Dysfunction in ankyrin-B-dependent ion channel and transporter targeting causes human sinus node disease. *Proc Natl Acad Sci USA* 2008;**105**:15617–15622.
6. Zhu YB, Luo JW, Jiang F, Liu G. Genetic analysis of sick sinus syndrome in a family harboring compound *CACNA1C* and *TTN* mutations. *Mol Med Rep* 2018; **17**:7073–7080.
7. Benson DW, Wang DW, Dyment M, Knilans TK, Fish FA, Strieper MJ, Rhodes TH, George AL. Jr., Congenital sick sinus syndrome caused by recessive mutations in the cardiac sodium channel gene (SCN5A). *J Clin Investig* 2003;**112**: 1019–1028.
8. Makita N, Sasaki K, Groenewegen WA, Yokota T, Yokoshiki H, Murakami T, Tsutsui H. Congenital atrial standstill associated with coinheritance of a novel SCN5A mutation and connexin 40 polymorphisms. *Heart Rhythm* 2005;**2**: 1128–1134.
9. Makiyama T, Akao M, Tsuji K, Doi T, Ohno S, Takenaka K, Kobori A, Ninomiya T, Yoshida H, Takano M, Makita N, Yanagisawa F, Higashi Y, Takeyama Y, Kita T, Horie M. High risk for bradyarrhythmic complications in patients with Brugada syndrome caused by *SCN5A* gene mutations. *J Am Coll Cardiol* 2005;**46**: 2100–2106.
10. Ishikawa T, Ohno S, Murakami T, Yoshida K, Mishima H, Fukuoka T, Kimoto H, Sakamoto R, Ohkusa T, Aiba T, Nogami A, Sumitomo N, Shimizu W, Yoshiura KI, Horigome H, Horie M, Makita N. Sick sinus syndrome with *HCN4* mutations

Downloaded from https://academic.oup.com/eurheartj/advance-article/doi/10.1093/eurheartj/ehaa1108/6134552 by Medicinsk Bibliotek, Aalborg Sygehus SYD user on 11 March 2021

shows early onset and frequent association with atrial fibrillation and left ventricular noncompaction. *Heart Rhythm* 2017;**14**:717–724.

11. Milanesi R, Baruscotti M, Gnecchi-Ruscone T, DiFrancesco D. Familial sinus bradycardia associated with a mutation in the cardiac pacemaker channel. *N Engl J Med* 2006;**354**:151–157.
12. Nof E, Luria D, Brass D, Marek D, Lahat H, Reznik-Wolf H, Pras E, Dascal N, Eldar M, Glikson M. Point mutation in the *HCN4* cardiac ion channel pore affecting synthesis, trafficking, and functional expression is associated with familial asymptomatic sinus bradycardia. *Circulation* 2007;**116**:463–470.
13. Stallmeyer B, Kuß J, Kotthoff S, Zumhagen S, Vowinkel K, Rinné S, Matschke LA, Friedrich C, Schulze-Bahr E, Rust S, Seebohm G, Decher N, Schulze-Bahr E. A mutation in the G-protein gene *GNB2* causes familial sinus node and atrioventricular conduction dysfunction. *Circ Res* 2017;**120**:e33–e44.
14. Lodder EM, De Nittis P, Koopman CD, Wiszniewski W, Moura de Souza CF, Lahrouchi N, Guex N, Napolioni V, Tessadori F, Beekman L, Nannenberg EA, Boualla L, Blom NA, de Graaff W, Kamermans M, Cocciadiferro D, Malerba N, Mandriani B, Akdemir ZHC, Fish RJ, Eldomery MK, Ratbi I, Wilde AAM, de Boer T, Simonds WF, Neerman-Arbez M, Sutton VR, Kok F, Lupski JR, Reymond A, Bezzina CR, Bakkers J, Merla G. *GNB5* mutations cause an autosomal-recessive multisystem syndrome with sinus bradycardia and cognitive disability. *Am J Hum Genet* 2016;**99**:704–710.
15. Kuß J, Stallmeyer B, Goldstein M, Rinné S, Pees C, Zumhagen S, Seebohm G, Decher N, Pott L, Kienitz MC, Schulze-Bahr E. Familial sinus node disease caused by a gain of GIRK (G-protein activated inwardly rectifying K(+) channel) channel function. *Circ Genom Precis Med* 2019;**12**:e002238.
16. Holm H, Gudbjartsson DF, Sulem P, Masson G, Helgadottir HT, Zanon C, Magnusson OT, Helgason A, Saemundsdottir J, Gylfason A, Stefansdottir H, Gretarsdottir S, Matthiasson SE, Thorgeirsson GM, Jonasdottir A, Sigurdsson A, Stefansson H, Werge T, Rafnar T, Kiemeney LA, Parvez B, Muhammad R, Roden DM, Darbar D, Thorleifsson G, Walters GB, Kong A, Thorsteinsdottir U, Arnar DO, Stefansson K. A rare variant in *MYH6* is associated with high risk of sick sinus syndrome. *Nat Genet* 2011;**43**:316–320.
17. Jensen PN, Gronroos NN, Chen LY, Folsom AR, deFilippi C, Heckbert SR, Alonso A. Incidence of and risk factors for sick sinus syndrome in the general population. *J Am Coll Cardiol* 2014;**64**:531–538.
18. Demoulin JC, Kulbertus HE. Histopathological correlates of sinoatrial disease. *Br Heart J* 1978;**40**:1384–1389.
19. Lien WP, Lee YS, Chang FZ, Lee SY, Chen CM, Tsai HC. The sick sinus syndrome: natural history of dysfunction of the sinoatrial node. *Chest* 1977;**72**: 628–634.
20. Morris GM, Kalman JM. Fibrosis, electrics and genetics. Perspectives in sinoatrial node disease. *Circ J* 2014;**78**:1272–1282.
21. Gomes JA, Kang PS, Matheson M, Gough WB, Jr., El-Sherif N. Coexistence of sick sinus rhythm and atrial flutter-fibrillation. *Circulation* 1981;**63**:80–86.
22. Kezerashvili A, Krumerman AK, Fisher JD. Sinus node dysfunction in atrial fibrillation: cause or effect? *J Atr Fibrillation* 2008;**1**:30.
23. Thorolfsdottir RB, Sveinbjornsson G, Sulem P, Helgadottir A, Gretarsdottir S, Benonisdottir S, Magnusdottir A, Davidsson OB, Rajamani S, Roden DM, Darbar D, Pedersen TR, Sabatine MS, Jonsdottir I, Arnar DO, Thorsteinsdottir U, Gudbjartsson DF, Holm H, Stefansson K. A missense variant in *PLEC* increases risk of atrial fibrillation. *J Am Coll Cardiol* 2017;**70**:2157–2168.
24. Davey Smith G, Hemani G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Hum Mol Genet* 2014;**23**:R89–98.
25. Burgess S, Davey Smith G, Davies NM, Dudbridge F, Gill D, Glymour MM, Hartwig FP, Holmes MV, Minelli C, Relton CL, Theodoratou E. Guidelines for performing Mendelian randomization investigations. *Wellcome Open Res* 2020;**4**: 186.
26. Akoum N, McGann C, Vergara G, Badger T, Ranjan R, Mahnkopf C, Kholmovski E, Macleod R, Marrouche N. Atrial fibrosis quantified using late gadolinium enhancement MRI is associated with sinus node dysfunction requiring pacemaker implant. *J Cardiovasc Electrophysiol* 2012;**23**:44–50.
27. Chang HY, Lin YJ, Lo LW, Chang SL, Hu YF, Li CH, Chao TF, Yin WH, Chen SA. Sinus node dysfunction in atrial fibrillation patients: the evidence of regional atrial substrate remodelling. *Europace* 2013;**15**:205–211.
28. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J, Landray M, Liu B, Matthews P, Ong G, Pell J, Silman A, Young A, Sprosen T, Peakman T, Collins R. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* 2015;**12**:e1001779.
29. Krokstad S, Langhammer A, Hveem K, Holmen TL, Midthjell K, Stene TR, Bratberg G, Heggland J, Holmen J. Cohort profile: the HUNT study, Norway. *Int J Epidemiol* 2013;**42**:968–977.
30. Hansen TF, Banasik K, Erikstrup C, Pedersen OB, Westergaard D, Chmura PJ, Nielsen K, Thorner L, Hjalgrim H, Paarup H, Larsen MAH, Petersen M, Jennum P, Andersen S, Nyegaard M, Jemec GBE, Olesen J, Werge T, Johansson PI, Sorensen E, Brunak S, Ullum H, Burgdorf KS. DBDS genomic cohort, a prospective and comprehensive resource for integrative and temporal analysis of genetic, environmental and lifestyle factors affecting health of blood donors. *BMJ Open* 2019;**9**:e028401.
31. Gudbjartsson DF, Helgason H, Gudjonsson SA, Zink F, Oddson A, Gylfason A, Besenbacher S, Magnusson G, Halldorsson BV, Hjartarson E, Sigurdsson GT, Stacey SN, Frigge ML, Holm H, Saemundsdottir J, Helgadottir HT, Johannsdottir H, Sigfusson G, Thorgeirsson G, Sverrisson JT, Gretarsdottir S, Walters GB, Rafnar T, Thjodleifsson B, Bjornsson ES, Olafsson S, Thorarinsdottir H, Steingrimsdottir T, Gudmundsdottir TS, Theodors A, Jonasson JG, Sigurdsson A, Bjornsdottir G, Jonsson JJ, Thorarensen O, Ludvigsson P, Gudbjartsson H, Eyjolfsson GI, Sigurdardottir O, Olafsson I, Arnar DO, Magnusson OT, Kong A, Masson G, Thorsteinsdottir U, Helgason A, Sulem P, Stefansson K. Large-scale whole-genome sequencing of the Icelandic population. *Nat Genet* 2015;**47**: 435–444.
32. Evangelou E, Warren HR, Mosen-Ansorena D, Mifsud B, Pazoki R, Gao H, Ntritsos G, Dimou N, Cabrera CP, Karaman I, Ng FL, Evangelou M, Witkowska K, Tzanis E, Hellwege JN, Giri A, Velez Edwards DR, Sun YV, Cho K, Gaziano JM, Wilson PWF, Tsao PS, Kovesdy CP, Esko T, Magi R, Milani L, Almgren P, Boutin T, Debette S, Ding J, Giulianini F, Holliday EG, Jackson AU, Li-Gao R, Lin WY, Luan J, Mangino M, Oldmeadow C, Prins BP, Qian Y, Sargurupremraj M, Shah N, Surendran P, Theriault S, Verweij N, Willems SM, Zhao JH, Amouyel P, Connell J, de Mutsert R, Doney ASF, Farrall M, Menni C, Morris AD, Noordam R, Pare G, Poulter NR, Shields DC, Stanton A, Thom S, Abecasis G, Amin N, Arking DE, Ayers KL, Barbieri CM, Batini C, Bis JC, Blake T, Bochud M, Boehnke M, Boerwinkle E, Boomsma DI, Bottinger EP, Braund PS, Brumat M, Campbell A, Campbell H, Chakravarti A, Chambers JC, Chauhan G, Ciullo M, Cocca M, Collins F, Cordell HJ, Davies G, de Borst MH, de Geus EJ, Deary IJ, Deelen J, Del Greco MF, Demirkale CY, Dorr M, Ehret GB, Elosua R, Enroth S, Erzurumluoglu AM, Ferreira T, Franberg M, Franco OH, Gandin I, Gasparini P, Giedraitis V, Gieger C, Girotto G, Goel A, Gow AJ, Gudnason V, Guo X, Gyllensten U, Hamsten A, Harris TB, Harris SE, Hartman CA, Havulinna AS, Hicks AA, Hofer E, Hofman A, Hottenga JJ, Huffman JE, Hwang SJ, Ingelsson E, James A, Jansen R, Jarvelin MR, Joehanes R, Johansson A, Johnson AD, Joshi PK, Jousilahti P, Jukema JW, Jula A, Kahonen M, Kathiresan S, Keavney BD, Khaw KT, Knekt P, Knight J, Kolcic I, Kooner JS, Koskinen S, Kristiansson K, Kutalik Z, Laan M, Larson M, Launer LJ, Lehne B, Lehtimaki T, Liewald DCM, Lin L, Lind L, Lindgren CM, Liu Y, Loos RJF, Lopez LM, Lu Y, Lyytikainen LP, Mahajan A, Mamasoula C, Marrugat J, Marten J, Milaneschi Y, Morgan A, Morris AP, Morrison AC, Munson PJ, Nalls MA, Nandakumar P, Nelson CP, Niiranen T, Nolte IM, Nutile T, Oldehinkel AJ, Oostra BA, O'Reilly PF, Org E, Padmanabhan S, Palmas W, Palotie A, Pattie A, Penninx B, Perola M, Peters A, Polasek O, Pramstaller PP, Nguyen QT, Raitakari OT, Ren M, Rettig R, Rice K, Ridker PM, Ried JS, Riese H, Ripatti S, Robino A, Rose LM, Rotter JI, Rudan I, Ruggiero D, Saba Y, Sala CF, Salomaa V, Samani NJ, Sarin AP, Schmidt R, Schmidt H, Shrine N, Siscovick D, Smith AV, Snieder H, Sober S, Sorice R, Starr JM, Stott DJ, Strachan DP, Strawbridge RJ, Sundstrom J, Swertz MA, Taylor KD, Teumer A, Tobin MD, Tomaszewski M, Toniolo D, Traglia M, Trompet S, Tuomilehto J, Tzourio C, Uitterlinden AG, Vaez A, van der Most PJ, van Duijn CM, Vergnaud AC, Verwoert GC, Vitart V, Volker U, Vollenweider P, Vuckovic D, Watkins H, Wild SH, Willemsen G, Wilson JF, Wright AF, Yao J, Zemunik T, Zhang W, Attia JR, Butterworth AS, Chasman DI, Conen D, Cucca F, Danesh J, Hayward C, Howson JMM, Laakso M, Lakatta EG, Langenberg C, Melander O, Mook-Kanamori DO, Palmer CNA, Risch L, Scott RA, Scott RJ, Sever P, Spector TD, van der Harst P, Wareham NJ, Zeggini E, Levy D, Munroe PB, Newton-Cheh C, Brown MJ, Metspalu A, Hung AM, O'Donnell CJ, Edwards TL, Psaty BM, Tzoulaki I, Barnes MR, Wain LV, Elliott P, Caulfield MJ. Genetic analysis of over 1 million people identifies 535 new loci associated with blood pressure traits. *Nat Genet* 2018;**50**:1412–1425.
33. Welsh S, Peakman T, Sheard S, Almond R. Comparison of DNA quantification methodology used in the DNA extraction protocol for the UK Biobank cohort. *BMC Genomics* 2017;**18**:26.
34. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, Motyer A, Vukcevic D, Delaneau O, O'Connell J, Cortes A, Welsh S, Young A, Effingham M, McVean G, Leslie S, Donnelly P, Marchini J. The UK Biobank resource with deep phenotyping and genomic data. *Nature* 2018;**562**:203–209 .
35. Walter K, Min JL, Huang J, Crooks L, Memari Y, McCarthy S, Perry JR, Xu C, Futema M, Lawson D, Iotchkova V, Schiffels S, Hendricks AE, Danecek P, Li R, Floyd J, Wain LV, Barroso I, Humphries SE, Hurles ME, Zeggini E, Barrett JC, Plagnol V, Richards JB, Greenwood CM, Timpson NJ, Durbin R, Soranzo N; UK10K Consortium. The UK10K project identifies rare variants in health and disease. *Nature* 2015;**526**:82–90.
36. Sveinbjornsson G, Albrechtsen A, Zink F, Gudjonsson SA, Oddson A, Masson G, Holm H, Kong A, Thorsteinsdottir U, Sulem P, Gudbjartsson DF, Stefansson K. Weighting sequence variants based on their annotation increases power of whole-genome association studies. *Nat Genet* 2016;**48**:314–317.

Downloaded from https://academic.oup.com/eurheartj/advance-article/doi/10.1093/eurheartj/ehaa1108/6134552 by Medicinsk Bibliotek, Aalborg Sygehus SYD user on 11 March 2021

37. Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J, Patterson N, Daly MJ, Price AL, Neale BM; Schizophrenia Working Group of the Psychiatric Genomics Consortium. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* 2015;**47**:291–295.
38. Richardson TG, Harrison S, Hemani G, Davey Smith G. An atlas of polygenic risk score associations to highlight putative causal relationships across the human phenome. *Elife* 2019;**8**:e43657.
39. Stearns FW. One hundred years of pleiotropy: a retrospective. *Genetics* 2010; **186**:767–773.
40. Burgess S, Thompson SG. Interpreting findings from Mendelian randomization using the MR-Egger method. *Eur J Epidemiol* 2017;**32**:377–389.
41. Nielsen JB, Thorolfsdottir RB, Fritsche LG, Zhou W, Skov MW, Graham SE, Herron TJ, McCarthy S, Schmidt EM, Sveinbjornsson G, Surakka I, Mathis MR, Yamazaki M, Crawford RD, Gabrielsen ME, Skogholt AH, Holmen OL, Lin M, Wolford BN, Dey R, Dalen H, Sulem P, Chung JH, Backman JD, Arnar DO, Thorsteinsdottir U, Baras A, O'Dushlaine C, Holst AG, Wen X, Hornsby W, Dewey FE, Boehnke M, Kheterpal S, Mukherjee B, Lee S, Kang HM, Holm H, Kitzman J, Shavit JA, Jalife J, Brummett CM, Teslovich TM, Carey DJ, Gudbjartsson DF, Stefansson K, Abecasis GR, Hveem K, Willer CJ. Biobank-driven genomic discovery yields new insight into atrial fibrillation biology. *Nat Genet* 2018;**50**:1234–1239.
42. Roselli C, Chaffin MD, Weng LC, Aeschbacher S, Ahlberg G, Albert CM, Almgren P, Alonso A, Anderson CD, Aragam KG, Arking DE, Barnard J, Bartz TM, Benjamin EJ, Bihlmeyer NA, Bis JC, Bloom HL, Boerwinkle E, Bottinger EB, Brody JA, Calkins H, Campbell A, Cappola TP, Carlquist J, Chasman DI, Chen LY, Chen YI, Choi EK, Choi SH, Christophersen IE, Chung MK, Cole JW, Conen D, Cook J, Crijns HJ, Cutler MJ, Damrauer SM, Daniels BR, Darbar D, Delgado G, Denny JC, Dichgans M, Dorr M, Dudink EA, Dudley SC, Esa N, Esko T, Eskola M, Fatkin D, Felix SB, Ford I, Franco OH, Geelhoed B, Grewal RP, Gudnason V, Guo X, Gupta N, Gustafsson S, Gutmann R, Hamsten A, Harris TB, Hayward C, Heckbert SR, Hernesniemi J, Hocking LJ, Hofman A, Horimoto A, Huang J, Huang PL, Huffman J, Ingelsson E, Ipek EG, Ito K, Jimenez-Conde J, Johnson R, Jukema JW, Kaab S, Kahonen M, Kamatani Y, Kane JP, Kastrati A, Kathiresan S, Katschnig-Winter P, Kavousi M, Kessler T, Kietselaer BL, Kirchhof P, Kleber ME, Knight S, Krieger JE, Kubo M, Launer LJ, Laurikka J, Lehtimaki T, Leineweber K, Lemaitre RN, Li M, Lim HE, Lin HJ, Lin H, Lind L, Lindgren CM, Lokki ML, London B, Loos RJF, Low SK, Lu Y, Lyytikainen LP, Macfarlane PW, Magnusson PK, Mahajan A, Malik R, Mansur AJ, Marcus GM, Margolin L, Margulies KB, Marz W, McManus DD, Melander O, Mohanty S, Montgomery JA, Morley MP, Morris AP, Muller-Nurasyid M, Natale A, Nazarian S, Neumann B, Newton-Cheh C, Niemeijer MN, Nikus K, Nilsson P, Noordam R, Oellers H, Olesen MS, Orho-Melander M, Padmanabhan S, Pak HN, Pare G, Pedersen NL, Pera J, Pereira A, Porteous D, Psaty BM, Pulit SL, Pullinger CR, Rader DJ, Refsgaard L, Ribases M, Ridker PM, Rienstra M, Risch L, Roden DM, Rosand J, Rosenberg MA, Rost N, Rotter JI, Saba S, Sandhu RK, Schnabel RB, Schramm K, Schunkert H, Schurman C, Scott SA, Seppala I, Shaffer C, Shah S, Shalaby AA, Shim J, Shoemaker MB, Siland JE, Sinisalo J, Sinner MF, Slowik A, Smith AV, Smith BH, Smith JG, Smith JD, Smith NL, Soliman EZ, Sotoodehnia N, Stricker BH, Sun A, Sun H, Svendsen JH, Tanaka T, Tanriverdi K, Taylor KD, Teder-Laving M, Teumer A, Theriault S, Trompet S, Tucker NR, Tveit A, Uitterlinden AG, Van Der Harst P, Van Gelder IC, Van Wagoner DR, Verweij N, Vlachopoulou E, Volker U, Wang B, Weeke PE, Weijs B, Weiss R, Weiss S, Wells QS, Wiggins KL, Wong JA, Woo D, Worrall BB, Yang PS, Yao J, Yoneda ZT, Zeller T, Zeng L, Lubitz SA, Lunetta KL, Ellinor PT. Multi-ethnic genome-wide association study for atrial fibrillation. *Nat Genet* 2018;**50**:1225–1233.
43. Guo Y, Chung W, Zhu Z, Shan Z, Li J, Liu S, Liang L. Genome-wide assessment for resting heart rate and shared genetics with cardiometabolic traits and type 2 diabetes. *J Am Coll Cardiol* 2019;**74**:2162–2174.
44. Yengo L, Sidorenko J, Kemper KE, Zheng Z, Wood AR, Weedon MN, Frayling TM, Hirschhorn J, Yang J, Visscher PM; the GIANT Consortium. Meta-analysis of genome-wide association studies for height and body mass index in approximately 700000 individuals of European ancestry. *Hum Mol Genet* 2018;**27**: 3641–3649.
45. van der Harst P, Verweij N. Identification of 64 novel genetic loci provides an expanded view on the genetic architecture of coronary artery disease. *Circ Res* 2018;**122**:433–443.
46. Malik R, Chauhan G, Traylor M, Sargurupremraj M, Okada Y, Mishra A, Rutten-Jacobs L, Giese A-K, van der Laan SW, Gretarsdottir S, Anderson CD, Chong M, Adams HHH, Ago T, Almgren P, Amouyel P, Ay H, Bartz TM, Benavente OR, Bevan S, Boncoraglio GB, Brown RD, Jr., Butterworth AS, Carrera C, Carty CL, Chasman DI, Chen W-M, Cole JW, Correa A, Cotlarciuc I, Cruchaga C, Danesh J, de Bakker PIW, DeStefano AL, den Hoed M, Duan Q, Engelter ST, Falcone GJ, Gottesman RF, Grewal RP, Gudnason V, Gustafsson S, Haessler J, Harris TB, Hassan A, Havulinna AS, Heckbert SR, Holliday EG, Howard G, Hsu F-C, Hyacinth HI, Ikram MA, Ingelsson E, Irvin MR, Jian X, Jiménez-Conde J, Johnson JA, Jukema JW, Kanai M, Keene KL, Kissela BM, Kleindorfer DO, Kooperberg C, Kubo M, Lange LA, Langefeld CD, Langenberg C, Launer LJ, Lee J-M, Lemmens R, Leys D, Lewis CM, Lin W-Y, Lindgren AG, Lorentzen E, Magnusson PK, Maguire J, Manichaikul A, McArdle PF, Meschia JF, Mitchell BD, Mosley TH, Nalls MA, Ninomiya T, O'Donnell MJ, Psaty BM, Pulit SL, Rannikmäe K, Reiner AP, Rexrode KM, Rice K, Rich SS, Ridker PM, Rost NS, Rothwell PM, Rotter JI, Rundek T, Sacco RL, Sakaue S, Sale MM, Salomaa V, Sapkota BR, Schmidt R, Schmidt CO, Schminke U, Sharma P, Slowik A, Sudlow CLM, Tanislav C, Tatlisumak T, Taylor KD, Thijs VNS, Thorleifsson G, Thorsteinsdottir U, Tiedt S, Trompet S, Tzourio C, van Duijn CM, Walters M, Wareham NJ, Wassertheil-Smoller S, Wilson JG, Wiggins KL, Yang Q, Yusuf S, Consortium AF, Cohorts for H, Aging Research in Genomic Epidemiology C, International Genomics of Blood Pressure C, Consortium I, Starnet, Bis JC, Pastinen T, Ruusalepp A, Schadt EE, Koplev S, Björkegren JLM, Codoni V, Civelek M, Smith NL, Trégouët DA, Christophersen IE, Roselli C, Lubitz SA, Ellinor PT, Tai ES, Kooner JS, Kato N, He J, van der Harst P, Elliott P, Chambers JC, Takeuchi F, Johnson AD, BioBank Japan Cooperative Hospital G, Consortium C, Consortium E-C, Consortium EP-I, International Stroke Genetics C, Consortium M, Neurology Working Group of the CC, Network NSG, Study UKYLD, Consortium M, Sanghera DK, Melander O, Jern C, Strbian D, Fernandez-Cadenas I, Longstreth WT, Jr., Rolfs A, Hata J, Woo D, Rosand J, Pare G, Hopewell JC, Saleheen D, Stefansson K, Worrall BB, Kittner SJ, Seshadri S, Fornage M, Markus HS, Howson JMM, Kamatani Y, Debette S, Dichgans M. Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes. *Nat Genet* 2018; **50**:524–537.
47. Shah S, Henry A, Roselli C, Lin H, Sveinbjörnsson G, Fatemifar G, Hedman ÅK, Wilk JB, Morley MP, Chaffin MD, Helgadottir A, Verweij N, Dehghan A, Almgren P, Andersson C, Aragam KG, Ärnlöv J, Backman JD, Biggs ML, Bloom HL, Brandimarto J, Brown MR, Buckbinder L, Carey DJ, Chasman DI, Chen X, Chen X, Chung J, Chutkow W, Cook JP, Delgado GE, Denaxas S, Doney AS, Dörr M, Dudley SC, Dunn ME, Engström G, Esko T, Felix SB, Finan C, Ford I, Ghanbari M, Ghasemi S, Giedraitis V, Giulianini F, Gottdiener JS, Gross S, Guðbjartsson DF, Gutmann R, Haggerty CM, van der Harst P, Hyde CL, Ingelsson E, Jukema JW, Kavousi M, Khaw K-T, Kleber ME, Køber L, Koekemoer A, Langenberg C, Lind L, Lindgren CM, London B, Lotta LA, Lovering RC, Luan Ja, Magnusson P, Mahajan A, Margulies KB, März W, Melander O, Mordi IR, Morgan T, Morris AD, Morris AP, Morrison AC, Nagle MW, Nelson CP, Niessner A, Niiranen T, O'Donoghue ML, Owens AT, Palmer CNA, Parry HM, Perola M, Portilla-Fernandez E, Psaty BM, Regeneron Genetics C, Rice KM, Ridker PM, Romaine SPR, Rotter JI, Salo P, Salomaa V, van Setten J, Shalaby AA, Smelser DT, Smith NL, Stender S, Stott DJ, Svensson P, Tammesoo M-L, Taylor KD, Teder-Laving M, Teumer A, Thorgeirsson G, Thorsteinsdottir U, Torp-Pedersen C, Trompet S, Tyl B, Uitterlinden AG, Veluchamy A, Völker U, Voors AA, Wang X, Wareham NJ, Waterworth D, Weeke PE, Weiss R, Wiggins KL, Xing H, Yerges-Armstrong LM, Yu B, Zannad F, Zhao JH, Hemingway H, Samani NJ, McMurray JJV, Yang J, Visscher PM, Newton-Cheh C, Malarstig A, Holm H, Lubitz SA, Sattar N, Holmes MV, Cappola TP, Asselbergs FW, Hingorani AD, Kuchenbaecker K, Ellinor PT, Lang CC, Stefansson K, Smith JG, Vasan RS, Swerdlow DI, Lumbers RT. Genome-wide association and Mendelian randomisation analysis provide insights into the pathogenesis of heart failure. *Nat Commun* 2020;**11**:163–163.
48. Ntalla I, Weng LC, Cartwright JH, Hall AW, Sveinbjornsson G, Tucker NR, Choi SH, Chaffin MD, Roselli C, Barnes MR, Mifsud B, Warren HR, Hayward C, Marten J, Cranley JJ, Concas MP, Gasparini P, Boutin T, Kolcic I, Polasek O, Rudan I, Araujo NM, Lima-Costa MF, Ribeiro ALP, Souza RP, Tarazona-Santos E, Giedraitis V, Ingelsson E, Mahajan A, Morris AP, Del Greco MF, Foco L, Gögele M, Hicks AA, Cook JP, Lind L, Lindgren CM, Sundström J, Nelson CP, Riaz MB, Samani NJ, Sinagra G, Ulivi S, Kähönen M, Mishra PP, Mononen N, Nikus K, Caulfield MJ, Dominiczak A, Padmanabhan S, Montasser ME, O'Connell JR, Ryan K, Shuldiner AR, Aeschbacher S, Conen D, Risch L, Thériault S, Hutri-Kähönen N, Lehtimäki T, Lyytikäinen LP, Raitakari OT, Barnes CLK, Campbell H, Joshi PK, Wilson JF, Isaacs A, Kors JA, van Duijn CM, Huang PL, Gudnason V, Harris TB, Launer LJ, Smith AV, Bottinger EP, Loos RJF, Nadkarni GN, Preuss MH, Correa A, Mei H, Wilson J, Meitinger T, Müller-Nurasyid M, Peters A, Waldenberger M, Mangino M, Spector TD, Rienstra M, van de Vegte YJ, van der Harst P, Verweij N, Kääb S, Schramm K, Sinner MF, Strauch K, Cutler MJ, Fatkin D, London B, Olesen M, Roden DM, Benjamin Shoemaker M, Gustav Smith J, Biggs ML, Bis JC, Brody JA, Psaty BM, Rice K, Sotoodehnia N, De Grandi A, Fuchsberger C, Pattaro C, Pramstaller PP, Ford I, Wouter Jukema J, Macfarlane PW, Trompet S, Dörr M, Felix SB, Völker U, Weiss S, Havulinna AS, Jula A, Sääksjärvi K, Salomaa V, Guo X, Heckbert SR, Lin HJ, Rotter JI, Taylor KD, Yao J, de Mutsert R, Maan AC, Mook-Kanamori DO, Noordam R, Cucca F, Ding J, Lakatta EG, Qian Y, Tarasov KV, Levy D, Lin H, Newton-Cheh CH, Lunetta KL, Murray AD, Porteous DJ, Smith BH, Stricker BH, Uitterlinden A, van den Berg ME, Haessler J, Jackson RD, Kooperberg C,

Downloaded from https://academic.oup.com/eurheartj/advance-article/doi/10.1093/eurheartj/ehaa1108/6134552 by Medicinsk Bibliotek, Aalborg Sygehus SYD user on 11 March 2021

Peters U, Reiner AP, Whitsel EA, Alonso A, Arking DE, Boerwinkle E, Ehret GB, Soliman EZ, Avery CL, Gogarten SM, Kerr KF, Laurie CC, Seyerle AA, Stilp A, Assa S, Abdullah Said M, Yldau van der Ende M, Lambiase PD, Orini M, Ramirez J, Van Duijvenboden S, Arnar DO, Gudbjartsson DF, Holm H, Sulem P, Thorleifsson G, Thorolfsdottir RB, Thorsteinsdottir U, Benjamin EJ, Tinker A, Stefansson K, Ellinor PT, Jamshidi Y, Lubitz SA, Munroe PB. Multi-ancestry GWAS of the electrocardiographic PR interval identifies 210 loci underlying cardiac conduction. *Nat Commun* 2020;11:2542.

49. van der Harst P, van Setten J, Verweij N, Vogler G, Franke L, Maurano MT, Wang X, Mateo Leach I, Eijgelsheim M, Sotoodehnia N, Hayward C, Sorice R, Meirelles O, Lyytikainen LP, Polasek O, Tanaka T, Arking DE, Ulivi S, Trompet S, Muller-Nurasyid M, Smith AV, Dorr M, Kerr KF, Magnani JW, Del Greco MF, Zhang W, Nolte IM, Silva CT, Padmanabhan S, Tragante V, Esko T, Abecasis GR, Adriaens ME, Andersen K, Barnett P, Bis JC, Bodmer R, Buckley BM, Campbell H, Cannon MV, Chakravarti A, Chen LY, Delitala A, Devereux RB, Doevendans PA, Dominiczak AF, Ferrucci L, Ford I, Gieger C, Harris TB, Haugen E, Heinig M, Hernandez DG, Hillege HL, Hirschhorn JN, Hofman A, Hubner N, Hwang SJ, Iorio A, Kahonen M, Kellis M, Kolcic I, Kooner IK, Kooner JS, Kors JA, Lakatta EG, Lage K, Launer LJ, Levy D, Lundby A, Macfarlane PW, May D, Meitinger T, Metspalu A, Nappo S, Naitza S, Neph S, Nord AS, Nutile T, Okin PM, Olsen JV, Oostra BA, Penninger JM, Pennacchio LA, Pers TH, Perz S, Peters A, Pinto YM, Pfeufer A, Pilia MG, Pramstaller PP, Prins BP, Raitakari OT, Raychaudhuri S, Rice KM, Rossin EJ, Rotter JI, Schafer S, Schlessinger D, Schmidt CO, Sehmi J, Sillje HHW, Sinagra G, Sinner MF, Slowikowski K, Soliman EZ, Spector TD, Spiering W, Stamatoyannopoulos JA, Stolk RP, Strauch K, Tan ST, Tarasov KV, Trinh B, Uitterlinden AG, van den Boogaard M, van Duijn CM, van Gilst WH, Viikari JS, Visscher PM, Vitart V, Volker U, Waldenberger M, Weichenberger CX, Westra HJ, Wijmenga C, Wolffenbuttel BH, Yang J, Bezzina CR, Munroe PB, Snieder H, Wright AF, Rudan I, Boyer LA, Asselbergs FW, van Veldhuisen DJ, Stricker BH, Psaty BM, Ciullo M, Sanna S, Lehtimaki T, Wilson JF, Bandinelli S, Alonso A, Gasparini P, Jukema JW, Kaab S, Gudnason V, Felix SB, Heckbert SR, de Boer RA, Newton-Cheh C, Hicks AA, Chambers JC, Jamshidi Y, Visel A, Christoffels VM, Isaacs A, Samani NJ, de Bakker PIW. 52 genetic loci influencing myocardial mass. *J Am Coll Cardiol* 2016;**68**(13):1435–1448.
50. Gudbjartsson DF, Arnar DO, Helgadottir A, Gretarsdottir S, Holm H, Sigurdsson A, Jonasdottir A, Baker A, Thorleifsson G, Kristjansson K, Palsson A, Blondal T, Sulem P, Backman VM, Hardarson GA, Palsdottir E, Helgason A, Sigurjonsdottir R, Sverrisson JT, Kostulas K, Ng MC, Baum L, So WY, Wong KS, Chan JC, Furie KL, Greenberg SM, Sale M, Kelly P, MacRae CA, Smith EE, Rosand J, Hillert J, Ma RC, Ellinor PT, Thorgeirsson G, Gulcher JR, Kong A, Thorsteinsdottir U, Stefansson K. Variants conferring risk of atrial fibrillation on chromosome 4q25. *Nature* 2007;**448**:353–357.
51. Gudbjartsson DF, Holm H, Gretarsdottir S, Thorleifsson G, Walters GB, Thorgeirsson G, Gulcher J, Mathiesen EB, Njolstad I, Nyrnes A, Wilsgaard T, Hald EM, Hveem K, Stoltenberg C, Kucera G, Stubblefield T, Carter S, Roden D, Ng MC, Baum L, So WY, Wong KS, Chan JC, Gieger C, Wichmann HE, Gschwendtner A, Dichgans M, Kuhlenbaumer G, Berger K, Ringelstein EB, Bevan S, Markus HS, Kostulas K, Hillert J, Sveinbjornsdottir S, Valdimarsson EM, Lochen ML, Ma RC, Darbar D, Kong A, Arnar DO, Thorsteinsdottir U, Stefansson K. A sequence variant in *ZFHX3* on 16q22 associates with atrial fibrillation and ischemic stroke. *Nat Genet* 2009;**41**:876–878.
52. Norland K, Sveinbjornsson G, Thorolfsdottir RB, Davidsson OB, Tragante V, Rajamani S, Helgadottir A, Gretarsdottir S, van Setten J, Asselbergs FW, Sverrisson JT, Stephensen SS, Oskarsson G, Sigurdsson EL, Andersen K, Danielsen R, Thorgeirsson G, Thorsteinsdottir U, Arnar DO, Sulem P, Holm H, Gudbjartsson DF, Stefansson K. Sequence variants with large effects on cardiac electrophysiology and disease. *Nat Commun* 2019;**10**:4803.
53. Pfeufer A, van Noord C, Marciante KD, Arking DE, Larson MG, Smith AV, Tarasov KV, Müller M, Sotoodehnia N, Sinner MF, Verwoert GC, Li M, Kao WHL, Köttgen A, Coresh J, Bis JC, Psaty BM, Rice K, Rotter JI, Rivadeneira F, Hofman A, Kors JA, Stricker BHC, Uitterlinden AG, van Duijn CM, Beckmann BM, Sauter W, Gieger C, Lubitz SA, Newton-Cheh C, Wang TJ, Magnani JW, Schnabel RB, Chung MK, Barnard J, Smith JD, Van Wagoner DR, Vasan RS, Aspelund T, Eiriksdottir G, Harris TB, Launer LJ, Najjar SS, Lakatta E, Schlessinger D, Uda M, Abecasis GR, Müller-Myhsok B, Ehret GB, Boerwinkle E, Chakravarti A, Soliman EZ, Lunetta KL, Perz S, Wichmann HE, Meitinger T, Levy D, Gudnason V, Ellinor PT, Sanna S, Kääb S, Witteman JCM, Alonso A, Benjamin EJ, Heckbert SR. Genome-wide association study of PR interval. *Nat Genet* 2010;**42**:153–159.
54. Holm H, Gudbjartsson DF, Arnar DO, Thorleifsson G, Thorgeirsson G, Stefansdottir H, Gudjonsson SA, Jonasdottir A, Mathiesen EB, Njolstad I, Nyrnes A, Wilsgaard T, Hald EM, Hveem K, Stoltenberg C, Lochen ML, Kong A, Thorsteinsdottir U, Stefansson K. Several common variants modulate heart rate, PR interval and QRS duration. *Nat Genet* 2010;**42**:117–122.
55. Human Genomics. The genotype-tissue expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* 2015;**348**:648–660.
56. Aalen OO, Johansen S. An empirical transition matrix for nonhomogeneous Markov-chains based on censored observations. *Scand Stat Theory Appl* 1978;**5**: 141–150.
57. Semelka M, Gera J, Usman S. Sick sinus syndrome: a review. *Am Fam Physician* 2013;**87**:691–696.
58. Ku NO, Gish R, Wright TL, Omary MB. Keratin 8 mutations in patients with cryptogenic liver disease. *N Engl J Med* 2001;**344**:1580–1587.
59. Benjamin EJ, Rice KM, Arking DE, Pfeufer A, van Noord C, Smith AV, Schnabel RB, Bis JC, Boerwinkle E, Sinner MF, Dehghan A, Lubitz SA, D'Agostino RB Sr, Lumley T, Ehret GB, Heeringa J, Aspelund T, Newton-Cheh C, Larson MG, Marciante KD, Soliman EZ, Rivadeneira F, Wang TJ, Eiríksdottir G, Levy D, Psaty BM, Li M, Chamberlain AM, Hofman A, Vasan RS, Harris TB, Rotter JI, Kao WHL, Agarwal SK, Stricker BHC, Wang K, Launer LJ, Smith NL, Chakravarti A, Uitterlinden AG, Wolf PA, Sotoodehnia N, Köttgen A, van Duijn CM, Meitinger T, Mueller M, Perz S, Steinbeck G, Wichmann H-E, Lunetta KL, Heckbert SR, Gudnason V, Alonso A, Kääb S, Ellinor PT, Witteman JCM. Variants in *ZFHX3* are associated with atrial fibrillation in individuals of European ancestry. *Nat Genet* 2009;**41**:879–881.
60. Nielsen JB, Fritsche LG, Zhou W, Teslovich TM, Holmen OL, Gustafsson S, Gabrielsen ME, Schmidt EM, Beaumont R, Wolford BN, Lin M, Brummett CM, Preuss MH, Refsgaard L, Bottinger EP, Graham SE, Surakka I, Chu Y, Skogholt AH, Dalen H, Boyle AP, Oral H, Herron TJ, Kitzman J, Jalife J, Svendsen JH, Olesen MS, Njølstad I, Løchen M-L, Baras A, Gottesman O, Marcketta A, O'Dushlaine C, Ritchie MD, Wilsgaard T, Loos RJF, Frayling TM, Boehnke M, Ingelsson E, Carey DJ, Dewey FE, Kang HM, Abecasis GR, Hveem K, Willer CJ. Genome-wide study of atrial fibrillation identifies seven risk loci and highlights biological pathways and regulatory elements involved in cardiac development. *Am J Hum Genet* 2018;**102**:103–115.
61. Ritchie MD, Denny JC, Zuvich RL, Crawford DC, Schildcrout JS, Bastarache L, Ramirez AH, Mosley JD, Pulley JM, Basford MA, Bradford Y, Rasmussen LV, Pathak J, Chute CG, Kullo IJ, McCarty CA, Chisholm RL, Kho AN, Carlson CS, Larson EB, Jarvik GP, Sotoodehnia N, Manolio TA, Li R, Masys DR, Haines JL, Roden DM. Genome- and phenome-wide analyses of cardiac conduction identifies markers of arrhythmia risk. *Circulation* 2013;**127**:1377–1385.
62. Bjornsson T, Thorolfsdottir RB, Sveinbjornsson G, Sulem P, Norddahl GL, Helgadottir A, Gretarsdottir S, Magnusdottir A, Danielsen R, Sigurdsson EL, Adalsteinsdottir B, Gunnarsson SI, Jonsdottir I, Arnar DO, Helgason H, Gudbjartsson T, Gudbjartsson DF, Thorsteinsdottir U, Holm H, Stefansson K. A rare missense mutation in MYH6 associates with non-syndromic coarctation of the aorta. *Eur Heart J* 2018;**39**:3243–3249.
63. Helgadottir A, Thorleifsson G, Gretarsdottir S, Stefansson OA, Tragante V, Thorolfsdottir RB, Jonsdottir I, Bjornsson T, Steinthorsdottir V, Verweij N, Nielsen JB, Zhou W, Folkersen L, Martinsson A, Heydarpour M, Prakash S, Oskarsson G, Gudbjartsson T, Geirsson A, Olafsson I, Sigurdsson EL, Almgren P, Melander O, Franco-Cereceda A, Hamsten A, Fritsche L, Lin M, Yang B, Hornsby W, Guo D, Brummett CM, Abecasis G, Mathis M, Milewicz D, Body SC, Eriksson P, Willer CJ, Hveem K, Newton-Cheh C, Smith JG, Danielsen R, Thorgeirsson G, Thorsteinsdottir U, Gudbjartsson DF, Holm H, Stefansson K. Genome-wide analysis yields new loci associating with aortic valve stenosis. *Nat Commun* 2018;**9**:987.
64. Ramírez J, Duijvenboden SV, Ntalla I, Mifsud B, Warren HR, Tzanis E, Orini M, Tinker A, Lambiase PD, Munroe PB. Thirty loci identified for heart rate response to exercise and recovery implicate autonomic nervous system. *Nat Commun* 2018;**9**:1947.
65. Verweij N, van de Vegte YJ, van der Harst P. Genetic study links components of the autonomous nervous system to heart-rate profile during exercise. *Nat Commun* 2018;**9**:898.
66. Rosenberg MA, Patton KK, Sotoodehnia N, Karas MG, Kizer JR, Zimetbaum PJ, Chang JD, Siscovick D, Gottdiener JS, Kronmal RA, Heckbert SR, Mukamal KJ. The impact of height on the risk of atrial fibrillation: the Cardiovascular Health Study. *Eur Heart J* 2012;**33**:2709–2717.
67. Schmidt M, Botker HE, Pedersen L, Sorensen HT. Adult height and risk of ischemic heart disease, atrial fibrillation, stroke, venous thromboembolism, and premature death: a population based 36-year follow-up study. *Eur J Epidemiol* 2014;**29**:111–118.
68. Levin M, Judy R, Gill D, Vujkovic M, Hyman M, Nazarian S, Rader D, Voight B, Damrauer S. Genetics of height and risk of atrial fibrillation: a Mendelian randomization study. PLoS Medicine 2019;**17**:e1003288.
69. Dobrzynski H, Boyett MR, Anderson RH. New insights into pacemaker activity: promoting understanding of sick sinus syndrome. *Circulation* 2007;**115**: 1921–1932.
70. Pan X, Hobbs RP, Coulombe PA. The expanding significance of keratin intermediate filaments in normal and diseased epithelia. *Curr Opin Cell Biol* 2013;**25**: 47–56.

Downloaded from https://academic.oup.com/eurheartj/advance-article/doi/10.1093/eurheartj/ehaa1108/6134552 by Medicinsk Bibliotek, Aalborg Sygehus SYD user on 11 March 2021

71. Coulombe PA, Omary MB. ‘Hard’ and ‘soft’ principles defining the structure, function and regulation of keratin intermediate filaments. *Curr Opin Cell Biol* 2002;**14**:110–122.
72. Moll R, Divo M, Langbein L. The human keratins: biology and pathology. *Histochem Cell Biol* 2008;**129**:705–733.
73. O’Neill A, Williams MW, Resneck WG, Milner DJ, Capetanaki Y, Bloch RJ. Sarcolemmal organization in skeletal muscle lacking desmin: evidence for cytokeratins associated with the membrane skeleton at costameres. *Mol Biol Cell* 2002;**13**:2347–2359.
74. Ursitti JA, Lee PC, Resneck WG, McNally MM, Bowman AL, O’Neill A, Stone MR, Bloch RJ. Cloning and characterization of cytokeratins 8 and 19 in adult rat striated muscle. Interaction with the dystrophin glycoprotein complex. *J Biol Chem* 2004;**279**:41830–41838.
75. Stone MR, O’Neill A, Catino D, Bloch RJ. Specific interaction of the actin-binding domain of dystrophin with intermediate filaments containing keratin 19. *Mol Biol Cell* 2005;**16**:4280–4293.
76. Papathanasiou S, Rickelt S, Soriano ME, Schips TG, Maier HJ, Davos CH, Varela A, Kaklamanis L, Mann DL, Capetanaki Y. Tumor necrosis factor-alpha confers cardioprotection through ectopic expression of keratins K8 and K18. *Nat Med* 2015;**21**:1076–1084.
77. Owens DW, Wilson NJ, Hill AJ, Rugg EL, Porter RM, Hutcheson AM, Quinlan RA, van Heel D, Parkes M, Jewell DP, Campbell SS, Ghosh S, Satsangi J, Lane EB. Human keratin 8 mutations that disturb filament assembly observed in inflammatory bowel disease patients. *J Cell Sci* 2004;**117**:1989–1999.
78. Ku NO, Omary MB. A disease- and phosphorylation-related nonmechanical function for keratin 8. *J Cell Biol* 2006;**174**:115–125.
79. Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglou S, Sidow A. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res* 2005;**15**:901–913.
80. England J, Loughna S. Heavy and light roles: myosin in the morphogenesis of the heart. *Cell Mol Life Sci* 2013;**70**:1221–1239.
81. Granzier HL, Labeit S. The giant protein titin: a major player in myocardial mechanics, signaling, and disease. *Circ Res* 2004;**94**:284–295.
82. Gudbjartsson DF, Holm H, Sulem P, Masson G, Oddsson A, Magnusson OT, Saemundsdottir J, Helgadottir HT, Helgason H, Johannsdottir H, Gretarsdottir S, Gudjonsson SA, Njolstad I, Lochen ML, Baum L, Ma RC, Sigfusson G, Kong A, Thorgeirsson G, Sverrisson JT, Thorsteinsdottir U, Stefansson K, Arnar DO. A frameshift deletion in the sarcomere gene *MYL4* causes early-onset familial atrial fibrillation. *Eur Heart J* 2017;**38**:27–34.
83. Thorolfsdottir RB, Sveinbjornsson G, Sulem P, Nielsen JB, Jonsson S, Halldorsson GH, Melsted P, Ivarsdottir EV, Davidsson OB, Kristjansson RP, Thorleifsson G, Helgadottir A, Gretarsdottir S, Norddahl G, Rajamani S, Torfason B, Valgardsson AS, Sverrisson JT, Tragante V, Holmen OL, Asselbergs FW, Roden DM, Darbar D, Pedersen TR, Sabatine MS, Willer CJ, Lochen ML, Halldorsson BV, Jonsdottir I, Hveem K, Arnar DO, Thorsteinsdottir U, Gudbjartsson DF, Holm H, Stefansson K. Coding variants in *RPL3L* and *MYZAP* increase risk of atrial fibrillation. *Commun Biol* 2018;**1**:68.
84. Miyata S, Minobe W, Bristow MR, Leinwand LA. Myosin heavy chain isoform expression in the failing and nonfailing human heart. *Circ Res* 2000;**86**:386–390.
85. James J, Martin L, Krenz M, Quatman C, Jones F, Klevitsky R, Gulick J, Robbins J. Forced expression of alpha-myosin heavy chain in the rabbit ventricle results in cardioprotection under cardiomyopathic conditions. *Circulation* 2005;**111**:2339–2346.
86. Cazorla O, Freiburg A, Helmes M, Centner T, McNabb M, Wu Y, Trombitás K, Labeit S, Granzier H. Differential expression of cardiac titin isoforms and modulation of cellular stiffness. *Circ Res* 2000;**86**:59–67.
87. Nagueh SF, Shah G, Wu Y, Torre-Amione G, King NM, Lahmers S, Witt CC, Becker K, Labeit S, Granzier HL. Altered titin expression, myocardial stiffness, and left ventricular function in patients with dilated cardiomyopathy. *Circulation* 2004;**110**:155–162.
88. Piedra ME, Icardo JM, Albajar M, Rodriguez-Rey JC, Ros MA. Pitx2 participates in the late phase of the pathway controlling left-right asymmetry. *Cell* 1998;**94**:319–324.
89. Berry FB, Miura Y, Mihara K, Kaspar P, Sakata N, Hashimoto-Tamaoki T, Tamaoki T. Positive and negative regulation of myogenic differentiation of C2C12 cells by isoforms of the multiple homeodomain zinc finger transcription factor ATBF1. *J Biol Chem* 2001;**276**:25057–25065.
90. Chen X, Yu L, Shi S, Jiang H, Huang C, Desai M, Li Y, Barajas-Martinez H, Hu D. Neuronal Nav1.8 channels as a novel therapeutic target of acute atrial fibrillation prevention. *J Am Heart Assoc* 2016;**5**:e004050.
91. Qi B, Wei Y, Chen S, Zhou G, Li H, Xu J, Ding Y, Lu X, Zhao L, Zhang F, Chen G, Zhao J, Liu S. Nav1.8 channels in ganglionated plexi modulate atrial fibrillation inducibility. *Cardiovasc Res* 2014;**102**:480–486.
92. Jabbari J, Olesen MS, Yuan L, Nielsen JB, Liang B, Macri V, Christophersen IE, Nielsen N, Sajadieh A, Ellinor PT, Grunnet M, Haunso S, Holst AG, Svendsen JH, Jespersen T; on behalf of LuCamp. Common and rare variants in *SCN10A* modulate the risk of atrial fibrillation. *Circ Cardiovasc Genet* 2015;**8**:64–73.
93. Katritsis DG, Pokushalov E, Romanov A, Giazitzoglou E, Siontis GC, Po SS, Camm AJ, Ioannidis JP. Autonomic denervation added to pulmonary vein isolation for paroxysmal atrial fibrillation: a randomized clinical trial. *J Am Coll Cardiol* 2013;**62**:2318–2325.
94. Nakagawa H, Scherlag BJ, Patterson E, Ikeda A, Lockwood D, Jackman WM. Pathophysiologic basis of autonomic ganglionated plexus ablation in patients with atrial fibrillation. *Heart Rhythm* 2009;**6**:S26–34.
95. Pokushalov E, Romanov A, Katritsis DG, Artyomenko S, Shirokova N, Karaskov A, Mittal S, Steinberg JS. Ganglionated plexus ablation vs linear ablation in patients undergoing pulmonary vein isolation for persistent/long-standing persistent atrial fibrillation: a randomized comparison. *Heart Rhythm* 2013;**10**:1280–1286.
96. Buckley U, Rajendran PS, Shivkumar K. Ganglionated plexus ablation for atrial fibrillation: just because we can, does that mean we should? *Heart Rhythm* 2017;**14**:133–134.
97. Lo LW, Scherlag BJ, Chang HY, Lin YJ, Chen SA, Po SS. Paradoxical long-term proarrhythmic effects after ablating the "head station" ganglionated plexi of the vagal innervation to the heart. *Heart Rhythm* 2013;**10**:751–757.
98. Osman F, Kundu S, Tuan J, Jeilan M, Stafford PJ, Ng GA, Ng GA. Ganglionic plexus ablation during pulmonary vein isolation—predisposing to ventricular arrhythmias? *Indian Pacing Electrophysiol J* 2010;**10**:104–107.
99. Driessen AHG, Berger WR, Krul SPJ, van den Berg NWE, Neefs J, Piersma FR, Chan Pin Yin D, de Jong J, van Boven WP, de Groot JR. Ganglion plexus ablation in advanced atrial fibrillation: the AFACT study. *J Am Coll Cardiol* 2016;**68**:1155–1165.
100. Elvan A, Wylie K, Zipes DP. Pacing-induced chronic atrial fibrillation impairs sinus node function in dogs. Electrophysiological remodeling. *Circulation* 1996;**94**:2953–2960.
101. Mesubi OO, Anderson ME. Atrial remodelling in atrial fibrillation: CaMKII as a nodal proarrhythmic signal. *Cardiovasc Res* 2016;**109**:542–557.

Downloaded from https://academic.oup.com/eurheartj/advance-article/doi/10.1093/eurheartj/ehaa1108/6134552 by Medicinsk Bibliotek, Aalborg Sygehus SYD user on 11 March 2021

View publication stats

European Heart Journal (2023) **44**, 1927–1939
https://doi.org/10.1093/eurheartj/ehad142

TRANSLATIONAL RESEARCH
Genetics

Dyslipidemia, inflammation, calcification, and adiposity in aortic stenosis: a genome-wide study

Hao Yu Chen[1,2], Christian Dina[3], Aeron M. Small[6], Christian M. Shaffer[7], Rebecca T. Levinson[7], Anna Helgadóttir[8], Romain Capoulade[3], Hans Markus Munter[9], Andreas Martinsson[10,36], Benjamin J. Cairns[11], Linea C. Trudsø[20], Mary Hoekstra[1,2], Hannah A. Burr[1,2], Thomas W. Marsh[2,37], Scott M. Damrauer[12], Line Dufresne[2], Solena Le Scouarnec[3], David Messika-Zeitoun[4,5], Dilrini K. Ranatunga[13], Rachel A. Whitmer[14], Amélie Bonnefond[15,16], Garðar Sveinbjornsson[8], Ragnar Daníelsen[17], David O. Arnar[8,17,18], Gudmundur Thorgeirsson[8,18], Unnur Thorsteinsdottir[8,18], Daníel F. Gudbjartsson[8,19], Hilma Hólm[8], Jonas Ghouse[20], Morten Salling Olesen[20], Alex H. Christensen[20,21], Susan Mikkelsen[22], Rikke Louise Jacobsen[23], Joseph Dowsett[23], Ole Birger Vesterager Pedersen[25], Christian Erikstrup[22,26], Sisse R. Ostrowski[23,24], Regeneron Genetics Center[27], Christopher J. O'Donnell[28], Matthew J. Budoff[29], Vilmundur Gudnason[30], Wendy S. Post[31], Jerome I. Rotter[32], Mark Lathrop[9,37], Henning Bundgaard[33], Bengt Johansson[34], Johan Ljungberg[34], Ulf Näslund[34], Thierry Le Tourneau[3], J. Gustav Smith[10,35,36], Quinn S. Wells[7], Stefan Söderberg[34], Kári Stefánsson[8,18], Jean-Jacques Schott[3], Daniel J. Rader[38], Robert Clarke[11], James C. Engert[1,2,37]*†, and George Thanassoulis[1,2]*†, on behalf of the Therapeutic targets for AoRtic stenosis using GEneTics (TARGET) Consortium

[1]Division of Experimental Medicine, McGill University, 1001 Decarie Blvd., Room EM1.2218, Montreal, Quebec H4A 3J1, Canada; [2]Preventive and Genomic Cardiology, McGill University Health Centre and Research Institute, 1001 Decarie Blvd., Room D05.5120, Montreal, Quebec H4A 3J1, Canada; [3]Nantes Université, CHU Nantes, CNRS, INSERM, l'institut du thorax, 8 Quai Moncousu, Nantes F-44000, France; [4]Department of Cardiology, Assistance Publique - Hôpitaux de Paris, Bichat Hospital, Paris, France; [5]Division of Cardiology, University of Ottawa Heart Institute, Ottawa, Ontario, Canada; [6]Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, USA; [7]Vanderbilt Translational and Clinical Cardiovascular Research Center, Vanderbilt University Medical Center, Nashville, USA; [8]deCODE genetics/Amgen Inc., Reykjavik, Iceland; [9]McGill University and Genome Quebec Innovation Centre, Montreal, Canada; [10]Department of Cardiology, Clinical Sciences, Lund University, Sweden and Skåne University Hospital, Lund, Sweden; [11]MRC Population Health Research Unit, Clinical Trial Service Unit and Epidemiological Studies Unit, Nuffield Department of Population Health, University of Oxford, Oxford, UK; [12]Department of Surgery, Perelman School of Medicine, University of Pennsylvania, Philadelphia, USA; [13]Division of Research, Kaiser Permanente of Northern California, Oakland, USA; [14]Department of Public Health Sciences, University of California Davis, Davis, USA; [15]University Lille, Inserm, CNRS, CHU Lille, Institut Pasteur de Lille, UMR1283-8199 EGID, Lille, France; [16]Department of Metabolism, Imperial College London, London, UK; [17]Internal Medicine and Emergency Services, Landspitali—The National University Hospital of Iceland, Reykjavik, Iceland; [18]School of Health Sciences, Faculty of Medicine, University of Iceland, Reykjavik, Iceland; [19]School of Engineering and Natural Sciences, University of Iceland, Reykjavik, Iceland; [20]Laboratory for Molecular Cardiology, Department of Cardiology, Copenhagen University Hospital, Rigshospitalet, Copenhagen, Denmark; [21]Department of Cardiology, Herlev-Gentofte Hospital, Copenhagen, Denmark; [22]Department of Clinical Immunology, Aarhus University Hospital, Aarhus, Denmark; [23]Department of Clinical Immunology, Copenhagen University Hospital, Rigshospitalet, Copenhagen, Denmark; [24]Department of Clinical Medicine, University of Copenhagen, Copenhagen, Denmark; [25]Department of Clinical Immunology, Zealand University Hospital, Naestved, Denmark; [26]Department of Clinical Medicine, Aarhus University, Aarhus, Denmark; [27]Regeneron Genetics Center, New York, USA; [28]National Heart, Lung, and Blood Institute's and Boston University's Framingham Heart Study, Boston, USA; [29]The Lundquist Institute for Biomedical Innovation at Harbor-UCLA Medical Center, Torrance, USA; [30]Faculty of

---

* Corresponding author. Tel: +1 514 934 1934 ext. 35325, Fax: 514 933 6418, Email: jamie.engert@mcgill.ca (J.C.E.); Tel: +1 514 934 1934 ext. 35465, Fax: 514 843 2813, Email: george.thanassoulis@mcgill.ca (G.T.)

† These authors are co-senior authors.

© The Author(s) 2023. Published by Oxford University Press on behalf of the European Society of Cardiology.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from https://academic.oup.com/eurheartj/article/44/21/1927/7112365 by guest on 15 October 2025

Medicine, University of Iceland, Reykjavík, Iceland; [31]Division of Cardiology, Department of Medicine, The Johns Hopkins University School of Medicine, Baltimore, USA; [32]The Institute for Translational Genomics and Population Sciences, Department of Pediatrics, The Lundquist Institute for Biomedical Innovation at Harbor-UCLA Medical Center, Torrance, USA; [33]Department of Cardiology, Copenhagen University Hospital, Rigshospitalet, Copenhagen, Denmark; [34]Department of Public Health and Clinical Medicine, Umeå University, Umeå, Sweden; [35]Wallenberg Center for Molecular Medicine and Lund University Diabetes Center, Lund, Sweden; [36]The Wallenberg Laboratory/Department of Molecular and Clinical Medicine, Institute of Medicine, Gothenburg University and the Department of Cardiology, Sahlgrenska University Hospital, Gothenburg, Sweden; [37]Department of Human Genetics, McGill University, Montreal, Canada; and [38]Departments of Genetics and Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, USA

Received 2 May 2022; revised 20 January 2023; accepted 21 February 2023; online publish-ahead-of-print 11 April 2023

See the editorial comment for this article 'The journey towards identification of actionable molecular pathways in calcific aortic valve stenosis', by S. Thériault et al., https://doi.org/10.1093/eurheartj/ehad134.

Abstract

Aims Although highly heritable, the genetic etiology of calcific aortic stenosis (AS) remains incompletely understood. The aim of this study was to discover novel genetic contributors to AS and to integrate functional, expression, and cross-phenotype data to identify mechanisms of AS.

Methods and results A genome-wide meta-analysis of 11.6 million variants in 10 cohorts involving 653 867 European ancestry participants (13 765 cases) was performed. Seventeen loci were associated with AS at $P \leq 5 \times 10^{-8}$, of which 15 replicated in an independent cohort of 90 828 participants (7111 cases), including *CELSR2–SORT1*, *NLRP6*, and *SMC2*. A genetic risk score comprised of the index variants was associated with AS [odds ratio (OR) per standard deviation, 1.31; 95% confidence interval (CI), 1.26–1.35; $P = 2.7 \times 10^{-51}$] and aortic valve calcium (OR per standard deviation, 1.22; 95% CI, 1.08–1.37; $P = 1.4 \times 10^{-3}$), after adjustment for known risk factors. A phenome-wide association study indicated multiple associations with coronary artery disease, apolipoprotein B, and triglycerides. Mendelian randomization supported a causal role for apolipoprotein B-containing lipoprotein particles in AS (OR per g/L of apolipoprotein B, 3.85; 95% CI, 2.90–5.12; $P = 2.1 \times 10^{-20}$) and replicated previous findings of causality for lipoprotein(a) (OR per natural logarithm, 1.20; 95% CI, 1.17–1.23; $P = 4.8 \times 10^{-73}$) and body mass index (OR per kg/m^2, 1.07; 95% CI, 1.05–1.9; $P = 1.9 \times 10^{-12}$). Colocalization analyses using the GTEx database identified a role for differential expression of the genes *LPA*, *SORT1*, *ACTR2*, *NOTCH4*, *IL6R*, and *FADS*.

Conclusion Dyslipidemia, inflammation, calcification, and adiposity play important roles in the etiology of AS, implicating novel treatments and prevention strategies.

Downloaded from https://academic.oup.com/eurheartj/article/44/21/1927/7112365 by guest on 15 October 2025

Structured Graphical Abstract

Key Question

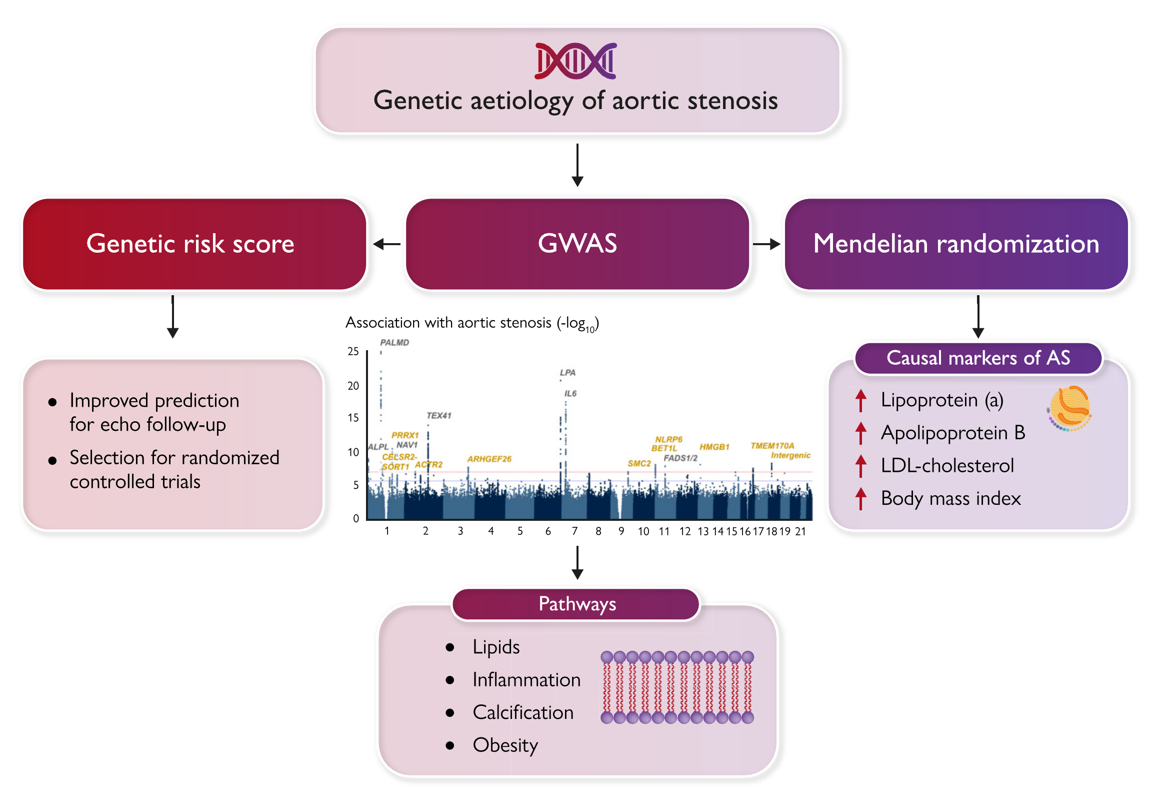
What genes and biological pathways play key roles in the aetiology of aortic stenosis (AS)?

Key Finding

Genetic variation that contributes to aortic stenosis was found to impact lipids, inflammation, calcification, and obesity.

Take Home Message

The identification of specific genes and pathways increases our understanding of the disease processes that result in aortic stenosis, and may eventually lead to the identification of high-risk patients and novel therapeutic approaches.


Genetic aetiology of aortic stenosis
Genetic risk score
GWAS
Mendelian randomization
Improved prediction for echo follow-up
Selection for randomized controlled trials
Association with aortic stenosis (-log10)
25
20
15
10
5
0
PALMD
LPA
IL6
TEX41
PRRX1
ALPL
NAV1
CELSR2-SORT1
ACTR2
ARHGEF26
NLRP6
BET1L
FADS1/2
HMGB1
SMC2
TMEM170A
Intergenic
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
21
Causal markers of AS
↑ Lipoprotein (a)
↑ Apolipoprotein B
↑ LDL-cholesterol
↑ Body mass index
Pathways
Lipids
Inflammation
Calcification
Obesity

Genetic etiology of aortic stenosis. This study meta-analyzed 13 765 AS cases vs. 640 102 controls and confirmed 15 genetic loci associated with AS. Downstream analyses implicated additional candidate genes involved in dyslipidemia, inflammation, calcification, and adiposity. The Manhattan plot shows variants with $P \geq 1 \times 10^{-25}$, for improved visualization. Genetic loci in grey were previously identified, and those in gold are new discoveries. Abbreviations: AS, aortic stenosis; GWAS, genome-wide association study; LDL, low-density lipoprotein.

Keywords Aortic stenosis • Genome-wide association study • Mendelian randomization • Phenome-wide association study • Gene expression • Genetic risk score

Translational perspective

This large-scale genetic study of calcific aortic stenosis (AS) in 653 867 European ancestry participants (13 765 cases) identified 15 robustly replicated genetic loci, including *SORT1–CELSR2*, involved in lipid metabolism, and *NLRP6*, involved in the inflammatory response. We provided evidence in favor of a causal association for apolipoprotein B, lipoprotein(a), body mass index, and low-density lipoprotein cholesterol. A genetic risk score including all identified loci was associated with both AS and aortic valve calcium and improved the classification of AS when added to risk factors. The differential expression of several genes in relevant tissues highlights their role in the etiology of AS. Together, these findings provide candidates for therapeutic targeting and highlight the use of a genetic risk score to improve clinical risk assessment.

Downloaded from https://academic.oup.com/eurheartj/article/44/21/1927/7112365 by guest on 15 October 2025

Introduction

Calcific aortic stenosis (AS) is the leading form of incident valvular heart disease in high-income populations.[1] In individuals over 75 years, the prevalence of AS is 10%–15% but is expected to more than double by 2040.[2] Although a replacement of the aortic valve is effective for severe AS cases,[3,4] there are no treatments to prevent progression to valve replacement. Furthermore, it remains unclear which patients are at high risk of a severe prognosis.

A genetic component contributes to the etiology of AS, as siblings of AS patients have more than four-fold the risk of AS.[5] Several families have also been identified with many affected members, including 1 extended family with 48 cases of severe AS.[6] Previous genome- and transcriptome-wide association efforts have identified seven genetic loci associated with AS,[7–11] including *LPA*[7] [which codes for the apolipoprotein(a) moiety of lipoprotein(a)], *IL6*[10] (which codes for interleukin-6), and the *FADS1/2* gene cluster[11] (which codes for desaturases involved in fatty acid biosynthesis). Additionally, Mendelian randomization studies supported a causal contribution of low-density lipoprotein cholesterol (LDL-C),[12] non-high-density lipoprotein cholesterol,[9] lipoprotein(a),[13] arachidonic acid[11], and body mass index (BMI)[14] to AS, suggesting susceptibility to AS is partially mediated by lipid metabolism and inflammation.

Identifying additional genetic loci for AS could provide novel targets for therapeutic intervention and improve risk stratification. Accordingly, we combined genome-wide association study (GWAS) results from 10 cohorts to identify novel loci. We conducted functional analyses for significant variants, examined their association with biomarkers and other diseases, assessed cross-ancestry transferability of variants, and developed an AS genetic risk score to assess its association with diagnosed AS. Finally, we investigated whether individual genes or gene sets were predicted to be differentially expressed in AS cases.

Methods

Cohorts and case definition

The cohorts included in the meta-analysis are listed in *Table 1*. A full description of the cohorts and their case definition for identifying AS is in the Supplementary material online.

Genome-wide meta-analysis for aortic stenosis

We performed centralized, cohort-specific quality control of genome-wide summary statistics for prevalent AS from 10 cohorts totaling 653 867 European ancestry participants (13 765 cases) (*Table 1*). With the exception of the Malmö Diet and Cancer Study, which excluded variants with >5% genotype missingness, minor allele frequency ≤1%, and Hardy–Weinberg equilibrium test $P \geq 1 \times 10^{-4}$, and deCODE, which excluded variants not found in the Haplotype Reference Consortium version r1.1 panel, quality control used unified criteria. For all cohorts, we included bi-allelic variants with non-ambiguous strands (i.e. no C/G or A/T allele pairs), imputation quality score ≥0.3, and minor allele frequency ≥0.001 and whose associations with AS had standard errors less than or equal to the median standard error plus five times the interquartile range, calculated from all summary statistics from each cohort. Using PLINK version 1.9, we performed inverse variance-weighted, fixed-effects meta-analysis for 11 591 806 variants with summary statistics that passed quality control and the allele frequency threshold in at least two cohorts. For independent ($r^2 \leq 0.01$) and genome-wide significant ($P \leq 5 \times 10^{-8}$) variants, i.e. index variants and variants in high linkage disequilibrium (LD) ($r^2 \geq 0.95$) in European ancestry individuals of the Genetic Epidemiology Research on Adult Health and Aging (GERA) cohort or the 1000 Genomes Project Phase 3,[15] we used the University of California Santa Cruz[16] GRCh37 assembly for genomic location.

Among 90 828 participants (7111 AS cases) from the Copenhagen Hospital Biobank or the Danish Blood Donor Study, we evaluated replication of the association with AS for all index variants. Associations were modeled using logistic regression adjusted for age, sex, and 10 principal components, with a $P \leq 0.05$ considered as replication. For all index variants, we also performed inverse variance-weighted, fixed-effects meta-analysis of summary statistics from the discovery and replication cohorts. For additional cohort details, see the Supplementary material online.

Region, gene-based, and functional analysis of variants

We used Annotate Variation (ANNOVAR)[17] to extract predicted function and pathogenicity of variants, including scores generated by Combined Annotation Dependent Depletion (CADD),[18] Deleterious Annotation of genetic variants using Neural Networks (DANN),[19] Linear INSIGHT (LINSIGHT),[20] Eigen-Principal Component (EIGEN-PC),[21] and Functional Analysis Through Hidden Markov Models-Multiple Kernel Learning (FATHMM-MKL) non-coding.[22] From Genotype-Tissue Expression Project (GTEx) version 8,[23] we identified significant expression quantitative trait loci in the aorta, left ventricle, liver, and whole blood. To identify gene regions associated with AS, we employed the single-nucleotide polymorphism (SNP)-wise mean approach of Multi-marker Analysis of GenoMic Annotation (MAGMA)[24] to test for association. We employed MetaXcan[25] and coloc[26] to identify variants whose effects on AS may be mediated by gene expression. We also applied the Genotype Imputed Gene Set Enrichment Analysis (GIGSEA)[27] approach, to assess whether sets of genes with shared function or regulation demonstrate differential predicted expression. We examined gene sets defined by the Kyoto Encyclopedia of Genes and Genomes (KEGG),[28] genes in the same pathway; Gene Ontology (GO),[29] genes with related functions; Functional ANnoTation Of the Mammalian genome version 5 (FANTOM5),[30] genes with shared transcription factor binding sites; and miRBase,[31] genes with the same microRNA seed sequence in their 3' untranslated region.

Genetic risk scores

We constructed three separate genetic risk scores. The first two used PLINK version 2.0[32] with weights estimated from a meta-analysis excluding the UK Biobank: a GRS_{18}, using all 18 genome-wide significant index variants from the discovery analysis, as well as a GRS_{559} including the 559 variants at $P < 1 \times 10^{-4}$. We assessed the association of GRS_{18} and GRS_{559} with AS in 220 159 unrelated White British participants in the UK Biobank aged 55 years or older (3091 cases) (see Supplementary material online) using a logistic regression model adjusted for age[2] and sex and then further adjusted for diabetes, LDL-C, systolic blood pressure, smoking (ever/never), BMI, and coronary artery disease (CAD). The UK Biobank participants in these and subsequent analyses differed slightly from the discovery analysis as an updated version of the data became available. In addition, we performed a polygenic risk score (PRS) analysis with LDpred2[33] in 244 641 UK Biobank participants (3410 cases). We also assessed the association of all the risk scores with the presence of aortic valve calcium (AVC) in the Multi-Ethnic Study of Atherosclerosis (MESA), using logistic regression adjusted for age and sex and then fully adjusted for fasting glucose, LDL-C, systolic blood pressure, smoking (ever/never), BMI, and coronary artery calcium. The PLINK-derived risk scores used 17 and 550 variants as neither variants nor proxies were available for some SNPs in this dataset. For the GRS_{17} and GRS_{550} risk scores, 2440 unrelated European participants (381 cases of AVC >0) were analyzed. The PRS analysis with LDpred2 included 2205 unrelated European participants (355 cases with AVC >0) of MESA. The area under the receiver operating characteristic curve (AUC) for the null hypothesis of no risk score was compared to other models using DeLong's test for two correlated ROC curves from the pROC R package.

Downloaded from https://academic.oup.com/eurheartj/article/44/21/1927/7112365 by guest on 15 October 2025

Table 1 Cohorts in the genome-wide meta-analysis

Cohort	Country	Aortic stenosis definition	No. of cases	No. controls	Median age (Quartile 1, Quartile 3)	Genotyping array	Imputation reference panel	No. of variants
Vanderbilt University Biobank	USA	her	759	7555	68 (62, 75)	Illumina Multi-Ethnic Genotyping Array	HRC version r1.1	10 689 407
CAVS-France1	France	Echocardiography	1261	1305	75 (7079)[a]	Affymetrix Axiom Genome-Wide CEU-1 Array	HRC version r1.1	10 395 306
CAVS-France2	France	Echocardiography	1495	2707	77 (70, 83)[a]	Affymetrix Axiom Genome-Wide Precision Medicine Research Array	HRC version r1.1	10 031 533
CAVS-France3	France	Echocardiography	367	2519	74 (67,79)[a]	Affymetrix Axiom Genome-Wide Precision Medicine Research Array	HRC version r1.1	9 884 426
deCODE	Iceland	her	2464	351 068	77 (68, 83)[a]	Illumina Chips	Icelandic reference panel by deCODE Genetics	9 812 907
GERA	USA	EHR	3469	51 723	66 (61, 74)	Affymetrix Axiom Genome-Wide EUR	HRC version r1.1	10 578 354
Malmö Diet and Cancer Study	Sweden	her	464	4878	58 (53, 63)	Illumina Human Omni Express Exome BeadChip	HRC version r1.1	6 130 056
Penn Medicine Biobank	USA	EHR	1593	4550	71 (62, 79)	Illumina Quad Omni Genotyping Chip	HRC version r.1.1	11 016 108
UK Biobank[b]	UK	her	1675	213 361	63 (59, 66)	Affymetrix UK Biobank Axiom Array	HRC version r1.1 and UK10K + 1000 Genomes Project Phase 3	9 927 329
Umeå University	Sweden	Surgery	218	436	60 (60, 66)	Affymetrix UK Biobank Axiom Array r3	HRC version r1.1	8 810 694

[a] In cases only.

[b] Provided are the number of cases and controls in the genome-wide association study for aortic stenosis. For analyses performed using UK Biobank data for the polygenic risk scores, phenome-wide association study, and Mendelian randomization, an updated dataset consisting of 2213 cases and 255 018 controls was used. See the UK Biobank cohort description for more details.

Abbreviations: CAVS, calcific aortic valve stenosis; EHR, electronic health records; EUR, European; GERA, Genetic Epidemiology Research on Adult Health and Aging; HRC, Haplotype Reference Consortium.

Downloaded from https://academic.oup.com/eurheartj/article/44/21/1927/7112365 by guest on 15 October 2025

Cross-ancestry and cross-phenotype associations

In the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) consortium, we extracted summary statistics[7] for the association of the index variants with prevalent AVC in an inverse variance-weighted fixed-effects meta-analysis of three independent cohorts totaling 6942 European participants (2245 cases), adjusting for age and sex. Variants that were not available were replaced with one in high LD ($r^2 \geq 0.8$) when available.

To examine the association of index variants with AS in other ancestries, we analyzed 1917 African American participants (86 cases) and 3482 Latin American participants (159 cases) of the GERA cohort, adjusted for age[2] and sex.

We performed a phenome-wide association study of the index variants with 58 diseases, serum biomarkers, and physiological measurements, among 257 231 unrelated White British UK Biobank participants aged 55 years or older. Disease cases were identified using hospital diagnosis codes, procedure codes, and causes of death. Levels of alkaline phosphatase, C-reactive protein, gamma-glutamyl transferase, lipoprotein(a), and triglycerides were natural logarithm transformed. Logistic and linear regression models were adjusted for age[2] and sex, except for breast cancer (analyzed only in women). A false discovery rate correction of 5% was applied across phenotypes for each variant tested.

For six traits associated with multiple variants in the phenome-wide association study, we performed two-sample Mendelian randomization to assess the causal contribution to AS. We performed a GWAS for each trait in unrelated UK Biobank White British participants (up to 383 533 participants). We constructed a genetic instrument for each biomarker using genome-wide significant variants ($331 \leq n \leq 702$) which were independent ($r^2 \leq 0.01$) with imputation quality ≥0.3 and minor allele frequency ≥0.001. We used the R package MendelianRandomization version 0.4.2[34] to estimate the inverse variance-weighted association with AS, using summary statistics from our discovery meta-analysis. The summary statistics for AS were generated with meta-analysis results that did not include UK Biobank so that instruments and outcomes were from non-overlapping cohorts. In secondary analyses, we also applied the contamination mixture, penalized weighted median, and Egger approaches.

Correlation and conditional analyses

We estimated the genetic correlation of AS with 157 cardiovascular biomarkers, risk factors, and diseases using the LD score regression method[35] as implemented on LD Hub.[36] We selected GWAS or meta-analyses which had been performed in European populations and applied a 5% false discovery rate correction.

To identify additional variants associated with AS, we used the conditional and joint analysis method[37] from the Genome-wide Complex Trait Analysis (GCTA) software[38] to re-estimate the summary statistics from our genome-wide meta-analysis conditioned upon the index variants. Variants not genome-wide significant in the original meta-analysis but which (i) became genome-wide significant in the conditional analysis and (ii) were independent ($r^2 \leq 0.01$) would be deemed to be novel associations.

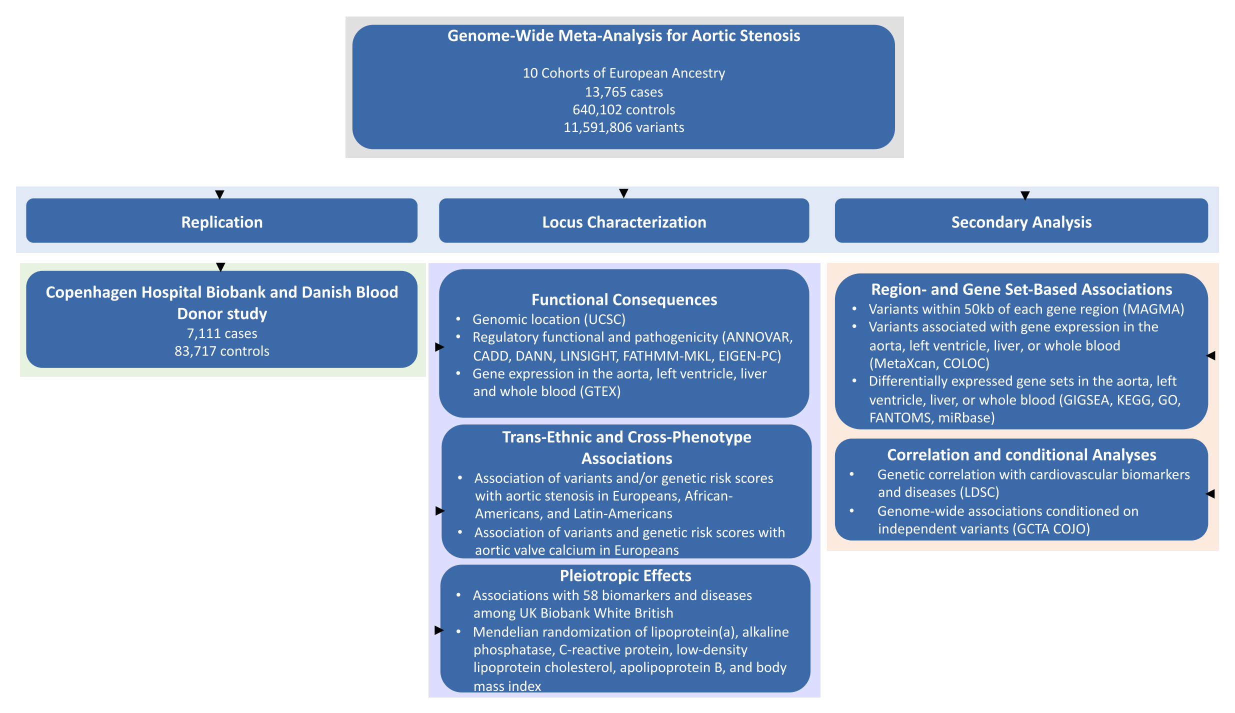

Genome-Wide Meta-Analysis for Aortic Stenosis
10 Cohorts of European Ancestry
13,765 cases
640,102 controls
11,591,806 variants
Replication
Locus Characterization
Secondary Analysis
Copenhagen Hospital Biobank and Danish Blood Donor study
7,111 cases
83,717 controls
Functional Consequences
• Genomic location (UCSC)
• Regulatory functional and pathogenicity (ANNOVAR, CADD, DANN, LINSIGHT, FATHMM-MKL, EIGEN-PC)
• Gene expression in the aorta, left ventricle, liver and whole blood (GTEX)
Trans-Ethnic and Cross-Phenotype Associations
• Association of variants and/or genetic risk scores with aortic stenosis in Europeans, African-Americans, and Latin-Americans
• Association of variants and genetic risk scores with aortic valve calcium in Europeans
Pleiotropic Effects
• Associations with 58 biomarkers and diseases among UK Biobank White British
• Mendelian randomization of lipoprotein(a), alkaline phosphatase, C-reactive protein, low-density lipoprotein cholesterol, apolipoprotein B, and body mass index
Region- and Gene Set-Based Associations
• Variants within 50kb of each gene region (MAGMA)
• Variants associated with gene expression in the aorta, left ventricle, liver, or whole blood (MetaXcan, COLOC)
• Differentially expressed gene sets in the aorta, left ventricle, liver, or whole blood (GIGSEA, KEGG, GO, FANTOMS, miRbase)
Correlation and conditional Analyses
• Genetic correlation with cardiovascular biomarkers and diseases (LDSC)
• Genome-wide associations conditioned on independent variants (GCTA COJO)

Figure 1 Design of the genome-wide meta-analysis and follow-up analyses. Abbreviations: ANNOVAR, Annotate Variation; CADD, Combined Annotation Dependent Depletion; DANN, Deleterious Annotation of genetic variants using Neural Networks; EIGEN-PC, Eigen-Principal Component; FANTOM5, Functional Annotation of the Mammalian Genome version 5; FATHMM-MKL, Functional Analysis Through Hidden Markov Models-Multiple Kernel Learning; GCTA COJO, Genome-wide Complex Trait Analysis Conditional and Joint Association Analysis; GIGSEA, Genotype Imputed Gene Set Enrichment Analysis; GO, Gene Ontology; GTEx, Genotype-Tissue Expression project; KEGG, Kyoto Encyclopedia of Genes and Genomics; LDSC, Linkage Disequilibrium Score Regression; LINSIGHT, Linear INSIGHT; MAGMA, Multi-marker Analysis of Genomic Annotation; UCSC, University of California Santa Cruz Genome Browser.

Downloaded from https://academic.oup.com/eurheartj/article/44/21/1927/7112365 by guest on 15 October 2025

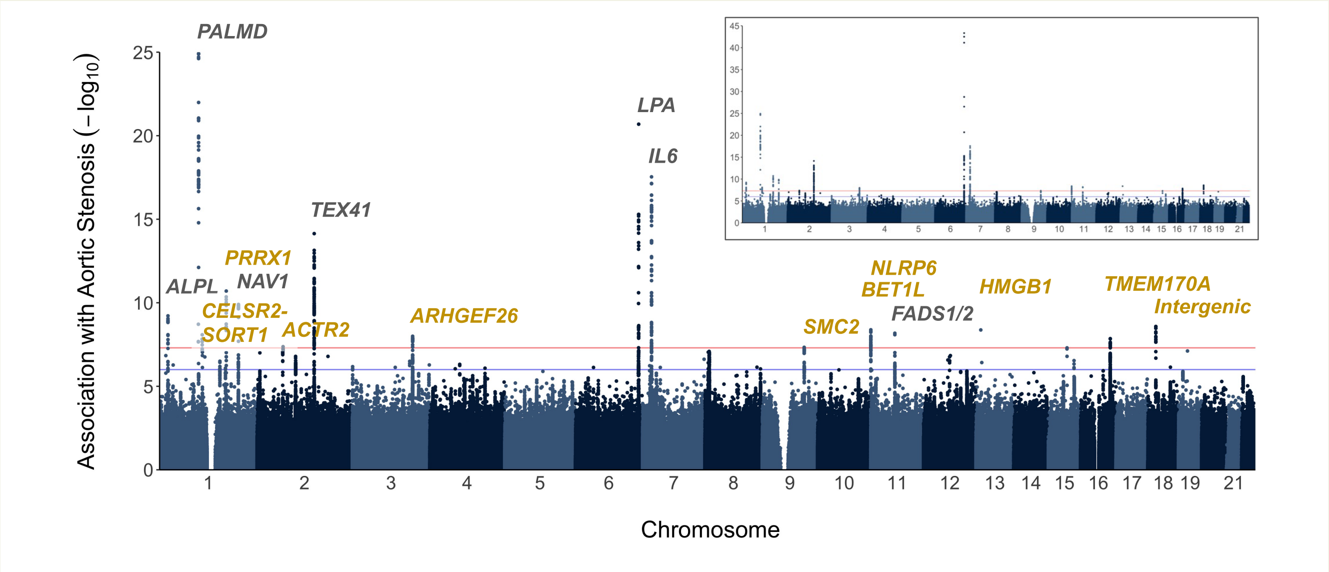


Figure 2 *P* for the association of 11 591 806 variants with aortic stenosis in the meta-analysis. The inset shows all associations, while the main plot shows variants with $P \geq 1 \times 10^{-25}$, for improved visualization. Genetic loci in grey were previously identified and those in gold are new discoveries.

Results

Genome-wide meta-analysis identifies 10 novel loci for aortic stenosis

We performed a genome-wide meta-analysis for AS using summary statistics from 10 European ancestry cohorts totaling 653 867 participants (13 765 cases) (*Table 1*). Estimates for each of 11 591 806 variants were combined in an inverse variance-weighted, fixed-effects model (see *Figure 1* for study design overview). We observed no evidence of strongly inflated test statistics in the meta-analysis (genomic inflation factor [λ] = 1.04 and LD score regression intercept = 1.020; Supplementary material online, *Figure S1*).

We identified 17 genetic loci containing one or more independent variants ($r^2 \leq 0.01$) which exceeded a genome-wide significance threshold ($P \leq 5 \times 10^{-8}$) for association with AS (*Figure 2*). We confirmed all 7 known loci and identified 10 loci not previously reported to be genome-wide significant. After pruning for variants in LD, we identified 18 independent variants ($r^2 \leq 0.01$) (see meta-analysis results in *Table 2* and forest plots in Supplementary material online, *Figure S2*). The association with AS of variants surrounding the index variants is provided in Supplementary material online, *Figure S3*).

Subsequently, we performed a replication study of the association for each of the 10 previously unreported loci with AS in 90 828 individuals (7111 AS cases) from the Danish Blood Donor Study and Copenhagen Hospital Biobank, and variants at 8 of these 10 loci were nominally associated with AS with concordant direction of effect ($P \leq 0.05$) (*Table 2*). The two least common variants, *BET1L* rs73386631 and *HMGB1* rs181753401, were not associated with AS in this cohort, and meta-analysis of the discovery and replication cohorts did not achieve genome-wide significance (*Table 2*).

When we re-estimated the association of variants with AS in the genome-wide meta-analysis, conditioned upon the 18 index variants, the *PLG* variant rs191108153 became genome-wide significant [odds ratio (OR) per T allele, 1.57; 95% CI, 1.34–1.83; $P = 9.6 \times 10^{-9}$]. Given the proximity of this variant to *LPA*, we tested the association of this variant with AS conditioned on genetically predicted levels of lipoprotein(a) and observed substantial attenuation (OR per T allele, 1.20; 95% CI, 1.03–1.40; $P = 0.019$).

We examined publicly available databases for functional effects of the index variants and their proxies in high LD ($r^2 \geq 0.95$) (see Supplementary material online, *Tables S1* and *S2*). *ARGHEF26* rs6794263 was in high LD ($r^2 = 0.99$)[39] with the missense variant rs13096373 (p.Phe203Ser). This substitution was predicted by the CADD software[40] to be in the 5% most deleterious substitutions (scaled C-score = 14.5). Variants in the *CELSR2–SORT1*, *PRRX1*, *NLRP6*, *PALMD*, and *IL6* loci and the intergenic region at 18q11.2 had high CADD scores, and variants at the *ACTR2* and *NLRP6* loci were in the 5% most pathogenic variants according to DANN[19] (ranked score ≥ 0.95). Variants at the *CELSR2–SORT1*, *PRRX1 PALMD*, *IL6*, *FADS1/2*, and 18q11.2 regions were classified as deleterious using the FATHMM-MKL non-coding approach ($P \geq 0.5$).[22]

Identified variants are also associated with aortic valve calcium

From a previous meta-analysis for AVC involving 6942 European participants from three cohorts in the CHARGE consortium (2245 participants with AVC >0),[7] we identified fixed-effects associations with prevalent AVC for the index variants. For unavailable variants, a proxy in high LD ($r^2 \geq 0.8$) was used, but no proxies were found for *LPA* rs140570886 and *HMGB1* rs181753401. Five variants were nominally associated with the presence of AVC in the same direction of effect as for AS ($P \leq 0.05$): *PRRX1* rs61817383 (OR per minor allele, 1.10; 95% CI, 1.01–1.21; $P = 0.035$), *ACTR2* rs62139062 (OR per minor allele, 1.11; 95% CI, 1.01–1.21; $P = 0.029$), *LPA* rs10455872 (OR per minor allele, 2.05; 95% CI, 1.63 to 2.57; $P = 9.0 \times 10^{-10}$), *NLRP6* rs17156153 (OR per minor allele, 1.17; 95% CI, 1.02–1.35; $P = 0.028$),

Downloaded from https://academic.oup.com/eurheartj/article/44/21/1927/7112365 by guest on 15 October 2025

Table 2 Association with aortic stenosis of new and previously identified genetic loci for aortic stenosis in the discovery and replication cohorts

Chr.	Locus	Variant	Minor allele	Discovery					Replication			Discovery and replication	
				MAF	Odds ratio per minor allele (95% CI)	*P*	I^2% (95% CI)	Imputation quality score range	MAF	Odds ratio per minor allele (95% CI)	*P*	odds ratio per minor allele (95% CI)	*P*
New loci													
1	*CELSR2–SORT1*	rs12740374	T	0.21	0.90 (0.87, 0.94)	8.4×10^{-09}	0 (0, 0)	0.77 to 1.00	0.23	0.95 (0.91, 0.99)	0.016	0.92 (0.90, 0.95)	2.4×10^{-09}
1	*PRRX1*	rs61817383	T	0.26	1.11 (1.08, 1.14)	2.0×10^{-11}	10 (0, 66)	0.94 to 1.00	0.26	1.06 (1.02, 1.10)	8.1×10^{-03}	1.09 (1.06, 1.12)	3.7×10^{-12}
2	*ACTR2*	rs62139062	T	0.28	1.09 (1.06, 1.12)	4.2×10^{-08}	32 (0, 68)	0.92 to 1.00	0.26	1.08 (1.04, 1.12)	2.3×10^{-04}	1.08 (1.06, 1.11)	4.4×10^{-11}
3	*ARHGEF26*	rs6794263	C	0.11	0.88 (0.84, 0.92)	1.0×10^{-08}	39 (0, 71)	0.98 to 1.00	0.11	0.93 (0.88, 0.99)	0.013	0.90 (0.87, 0.93)	1.5×10^{-09}
9	*SMC2*	rs55909255	C	0.39	1.08 (1.05, 1.11)	4.7×10^{-08}	53 (3, 77)	0.99 to 1.00	0.36	1.05 (1.01, 1.09)	0.014	1.07 (1.04, 1.09)	5.1×10^{-09}
11	*BET1L*	rs73386631	T	0.043	1.22 (1.14, 1.31)	2.8×10^{-08}	20 (0, 62)	0.37 to 1.00	0.038	1.06 (0.97, 1.17)	0.19	1.16 (1.10, 1.23)	1.9×10^{-07}
11	*NLRP6*	rs17156153	T	0.085	1.16 (1.10, 1.22)	4.2×10^{-09}	0 (0, 56)	0.73 to 1.00	0.075	1.16 (1.09, 1.24)	1.3×10^{-05}	1.16 (1.11, 1.21)	2.6×10^{-13}
13	*HMGB1*	rs181753401	A	2.4×10^{-03}	2.29 (1.74, 3.02)	4.2×10^{-09}	36 (0, 73)	0.40 to 1.00	3.9×10^{-03}	1.23 (0.92, 1.63)	0.16	1.69 (1.38, 2.06)	2.1×10^{-07}
16	*TMEM170A*	rs11643207	C	0.38	0.92 (0.89, 0.95)	1.4×10^{-08}	50 (0, 76)	0.98 to 1.00	0.39	0.92 (0.89, 0.96)	8.6×10^{-06}	0.92 (0.90, 0.94)	5.6×10^{-13}
18	*Intergenic* (18q11.2)	rs551520	T	0.23	0.91 (0.88, 0.94)	2.6×10^{-09}	27 (0, 65)	0.99 to 1.01	0.24	0.94 (0.90, 0.98)	2.8×10^{-03}	0.92 (0.90, 0.94)	6.8×10^{-11}
Previously identified loci													
1	*ALPL*	rs6696066	A	0.48	0.92 (0.89, 0.94)	6.1×10^{-10}	0 (0, 6.3)	0.99 to 1.00	0.47	0.93 (0.89, 0.96)	2.3×10^{-05}	0.92 (0.90, 0.94)	7.1×10^{-14}
1	*PALMD*	rs6702619	G	0.49	1.16 (1.13, 1.19)	1.2×10^{-25}	59 (18, 80)	0.97 to 1.00	0.51	1.17 (1.13, 1.21)	7.1×10^{-19}	1.16 (1.14, 1.19)	8.7×10^{-43}
1	*NAV1*	rs631556	A	0.41	1.10 (1.07, 1.13)	1.4×10^{-10}	0 (0, 44)	0.98 to 1.00	0.38	1.08 (1.04, 1.12)	5.3×10^{-05}	1.09 (1.07, 1.12)	4.1×10^{-14}
2	*TEX41*	rs7593336	G	0.38	1.12 (1.09, 1.15)	7.3×10^{-15}	59 (17, 79)	0.95 to 1.00	0.41	1.14 (1.10, 1.18)	2.7×10^{-12}	1.12 (1.10, 1.15)	1.7×10^{-25}
6	*LPA*	rs10455872	G	0.069	1.42 (1.35, 1.49)	4.6×10^{-44}	28 (0, 65)	0.76 to 1.00	0.081	1.50 (1.41, 1.60)	1.1×10^{-33}	1.45 (1.39, 1.51)	1.4×10^{-75}
		rs140570886	C	0.014	1.55 (1.40, 1.73)	5.1×10^{-16}	24 (0, 64)	0.83 to 1.00	0.010	1.31 (1.10, 1.56)	3.0×10^{-03}	1.48 (1.35, 1.62)	2.3×10^{-17}
7	*IL6*	rs1800797	A	0.45	1.13 (1.10, 1.16)	2.9×10^{-18}	45 (0, 73)	0.88 to 1.00	0.46	1.12 (1.08, 1.16)	9.3×10^{-11}	1.13 (1.10, 1.15)	1.9×10^{-27}
11	*FADS1/2*	rs174533	A	0.34	0.91 (0.88, 0.94)	6.7×10^{-09}	0 (0, 54)	0.93 to 1.00	0.34	0.93 (0.89, 0.96)	4.8×10^{-05}	0.92 (0.89, 0.94)	2.0×10^{-12}

Abbreviation: MAF, minor allele frequency.

Downloaded from https://academic.oup.com/eurheartj/article/44/21/1927/7112365 by guest on 15 October 2025

and *FADS1/2* rs174533 (OR per minor allele, 0.91; 95% CI, 0.83–0.99; $P = 0.024$) (see Supplementary material online, *Table S3*).

Genetic risk scores are associated with aortic stenosis and aortic valve calcium

The prevalence of AS increased across GRS_{18} risk score tertiles, ranging from 0.97% in the lowest tertile to 1.92% in the highest tertile (see Supplementary material online, *Figure S4*). Each 1 standard deviation (SD) higher GRS_{18} was associated with 37% higher odds of AS (OR per SD, 1.37; 95% CI, 1.32–1.41; $P = 3.0 \times 10^{-72}$) with an AUC of 0.697, when adjusted for age[2] and sex, with a similar OR after additional adjustment for diabetes, LDL-C, systolic blood pressure, smoking, BMI, and CAD (OR per SD, 1.31; 95% CI, 1.26–1.35; $P = 2.6 \times 10^{-51}$) (see Supplementary material online, *Table S4*). The AUC for this full model was 0.824, the addition of the genetic risk score modestly improving the AUC compared to a model restricted to non-genetic cardiovascular risk factors only (AUC = 0.818; $P_{\text{AUC difference}} = 5.9 \times 10^{-6}$). The GRS_{559} demonstrated stronger effect sizes and higher discrimination in the fully adjusted model (OR 1.53 per 1-SD GRS559, 95% CI 1.48–1.58; $P = 1.54 \times 10^{-141}$; AUC = 0.829; $P_{\text{AUC difference}} = 3.2 \times 10^{-9}$). In an additional sensitivity analysis, we dropped the six *LPA* region SNPs from the GRS_{559}, and this resulted in an OR of 1.50 per 1-SD for AS (CI 1.45–1.55; $P = 7.21 \times 10^{-130}$). Using the LDpred2 approach, we observed similar results with an OR of 1.45 per 1-SD (95% CI 1.41, 1.50; $P = 7.44 \times 10^{-122}$) and an AUC of 0.706 when adjusting only for age[2] and sex. In the model also adjusting for cardiovascular risk factors, we observed an OR of 1.40 (95% CI 1.36, 1.45; $P = 1.19 \times 10^{-84}$) and an AUC of 0.827 (see Supplementary material online, *Table S4*).

When we applied the GRS_{17} to MESA, the prevalence of AVC was 15.2%, 13.9%, and 17.7% in the first, second, and third tertiles, respectively (see Supplementary material online, *Figure S4*). After adjustment for age and sex, each 1 SD higher genetic risk score was associated with 22% higher odds of AVC (OR per SD, 1.22; 95% CI, 1.08–1.37; $P = 1.44 \times 10^{-3}$; AUC = 0.796), and this association persisted after additional adjustment for fasting glucose, LDL-C, systolic blood pressure, smoking, BMI, and coronary artery calcium (OR per SD, 1.23; 95% CI, 1.09–1.39; $P = 1.10 \times 10^{-3}$; AUC = 0.815). The GRS_{550} demonstrated weaker effects and worse discrimination with AVC when adjusted for age and sex (OR 1.15, 95% CI 1.02–1.29; $P = 0.022$; AUC = 0.794) as well as in the fully adjusted model (OR 1.15, 95% CI 1.02–1.30; $P = 0.023$; AUC = 0.814). Without the six *LPA* region SNPs, the OR was 1.11 per 1-SD for AVC (CI 0.98–1.25; $P = 0.093$). However, we observed stronger effects and better discrimination with LDpred2, an OR of 1.27 (1.12, 1.44; $P = 1.34 \times 10^{-9}$) and an AUC of 0.799 when adjusted for age and sex. In the fully adjusted model, we observed an OR of 1.32 (1.16, 1.50; $P = 2.24 \times 10^{-5}$) and an AUC of 0.823 (see Supplementary material online, *Table S4*).

A subset of the identified variants replicates in African and Latin Americans

When we examined the index variants among 1917 African American participants (86 cases) and 3482 Latin American participants (159 cases) from the GERA cohort, we observed replication ($P \leq 0.05$ for AS with concordant direction of effects) for *CELSR2–SORT1* rs12740374 in both ancestries, for *ALPL* rs6696066 and *NLRP6* rs17156153 in Latin Americans, and for *LPA* rs10455872 in African Americans (see Supplementary material online, *Table S5*).

Region-based analysis identifies additional associations with aortic stenosis

For each of 18 539 protein-coding genes, we used MAGMA[24] to examine the joint contribution of all variants in a gene region. We tested the mean association with AS of variants within 50 kb of each gene, correcting for LD between variants. We identified 95 regions associated with AS after Bonferroni correction (see Supplementary material online, *Table S6*), including regions spanning 11 of the 17 loci identified with single variant analysis (*ALPL*, *CELSR2*, *PRRX1*, *NAV1*, *ARHGEF26*, *LPA*, *IL6*, *BET1L*, *NLRP6*, *FADS1/2*, and *TMEM170A*). The *TMEM170A* gene region had a level of significance similar to an overlapping region defined by CFDP1 ($P = 2.9 \times 10^{-10}$ for *CFDP1* vs. $P = 4.5 \times 10^{-10}$ for *TMEM170A*). The three most significant regions not identified in the single variant-based approaches were *LDLR* ($P = 2.3 \times 10^{-10}$), *AGO2* ($P = 5.9 \times 10^{-10}$), and *XKR6* ($P = 9.8 \times 10^{-10}$). Although this approach did not account for LD between regions, these three regions were >100 Mb away from another association.

Expression-based analyses

We used MetaXcan[25] to analyze gene expression and expression quantitative trait loci (eQTL) extracted from GTEx project data.[23] We examined four tissues which may be involved in the AS disease process: aorta, left ventricle, liver, and whole blood. With a false discovery rate of 5%, we identified 42 genes with predicted expression that differed between cases and controls (see Supplementary material online, *Figure S5*). Increased *LPA* expression in the liver was predicted for AS cases (Z score = 5.25; $P = 1.5 \times 10^{-7}$). Expression of *IL6R* was inferred to be increased in the blood of cases (Z score = 4.66; $P = 3.1 \times 10^{-6}$). *NOTCH4* mRNA in the aorta was predicted to be decreased in cases (Z score = −4.07; $P = 4.7 \times 10^{-5}$). In contrast to these single tissue effects, inferred *RPS23* expression was lower in the blood, left ventricle, and aorta in AS cases ($P \leq 3.3 \times 10^{-5}$). The most significant, predicted differential expression was decreased *ZEB2* expression in the aorta of individuals with AS (Z score = −7.14; $P = 9.2 \times 10^{-13}$). Additional GIGSEA[27] were performed to infer the differential expression of genes (see Supplementary material online, *Table S7*).

As identified in GTEx,[23] the index variants at the *ACTR2* and *BET1L* loci were in high LD ($r^2 \geq 0.98$) with the most significant eQTL for their respective genes in the aorta. The index variant at the *ARHGEF26* locus was in high LD ($r^2 = 0.96$) with the most significant eQTL for *ARHGEF26* in the left ventricle. In addition, rs55909255 was associated with the expression of *SMC2* in the left ventricle, and rs140570886 was associated with the expression of *LPA* in the liver. Index variants and proxies at the *CELSR2–SORT1*, *IL6*, and *FADS1/2* loci were all eQTL for multiple genes in three to four tissues (see Supplementary material online, *Table S2*). Notably, the index variant at the *CELSR2–SORT1* locus was the most significant hepatic eQTL for *CELSR2* and *SORT1*, and the index variant at the *FADS1/2* locus was in high LD ($r^2 \geq 0.96$) with the most significant eQTL for *FADS1* in the aorta and liver. Neither the index variant at the *IL6* locus nor any variants in high LD ($r^2 \geq 0.95$) were associated with *IL6* expression in the tissues examined. However, the index variant was in high LD with the most significant eQTL for IL6 antisense RNA 1 (*IL6-AS1*) in the blood ($r^2 = 0.96$) (see Supplementary material online, *Table S2*). Several of these results were confirmed by additional colocalization analyses using MetaXcan and coloc (see Supplementary material online, *Figure S5*, and Supplementary material online, *Table S8*). Notably,

Downloaded from https://academic.oup.com/eurheartj/article/44/21/1927/7112365 by guest on 15 October 2025

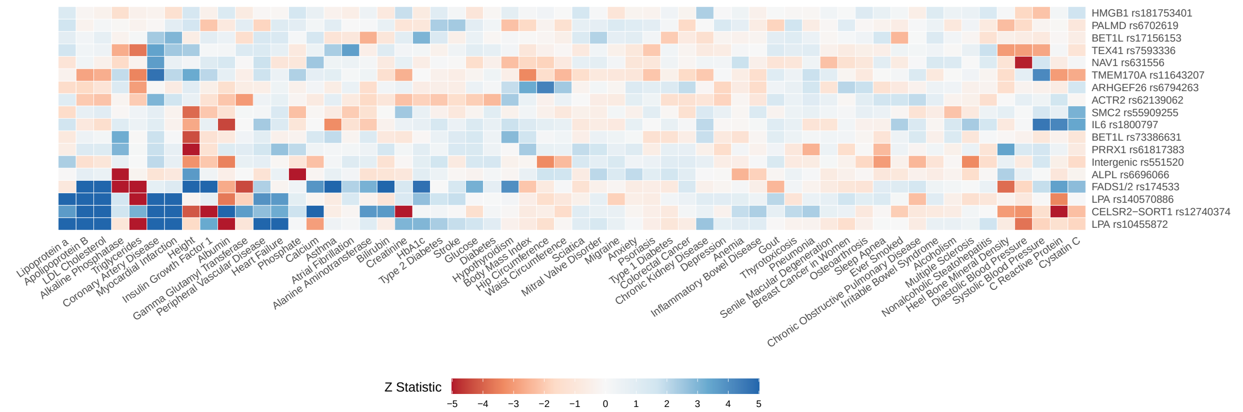


Figure 3 Association of aortic stenosis variants with biomarkers, physiological measurements, and diseases. Variants were ordered to reflect similarity in their associations with traits and vice versa. The strength and direction of associations are represented by cells of different colors, with blue cells indicating positive effects on the odds of aortic stenosis and red cells indicating inverse associations. For ease of visualization, Z statistics >5 or less than −5 were rounded to 5 and −5, respectively.

we identified high posterior probabilities for colocalization for SORT1 expression in the liver for rs6702619, as well as ACTR2 expression in the aorta for rs7556894 and FADS1 in whole blood and left ventricle for rs174535.

AS is genetically correlated with adiposity

Using LD score regression[35] as implemented on LD Hub,[36] we computed the genetic correlation of AS with 157 traits related to cardiovascular risk factors, metabolites, and immunological diseases using GWAS summary statistics. With a false discovery rate of 5%, we observed 32 (18.5%) traits that were genetically correlated with AS, most of which involved adiposity, glycemia, or lipids (see Supplementary material online, *Table S9*). Measures of adiposity were positively correlated with AS and represented seven of the 10 most significant correlations. The only inverse correlations were with high-density lipoprotein particles. The total lipid content of chylomicrons and large very low-density lipoprotein particles shared the highest absolute genetic correlation with AS (r_g, 0.31; 95% CI, 0.13 to 0.48; $P = 6.0 \times 10^{-4}$).

Pleiotropic effects of variants associated with aortic stenosis

We assessed the association of the index variants with 58 biomarkers, physiological measurements, and diseases among 257 231 UK Biobank White British participants. Traits were selected to reflect cardiovascular, cerebrovascular, immune, hepatic, and adiposity phenotypes. We observed 98 associations with false discovery rate-adjusted $P \leq 0.05$ and the five most significant associations involved apolipoprotein B, lipoprotein(a), LDL-C, or triglycerides (see Supplementary material online, *Table S10*). The traits with the greatest number of associations were height (eight variants); CAD, albumin, and triglycerides (six variants each); C-reactive protein and apolipoprotein B (five variants each); and heel bone mineral density, diastolic blood pressure, and LDL-C cholesterol (four variants each) (*Figure 3*; Supplementary material online, *Table S10*).

Blood-based biomarkers are causal contributors to aortic stenosis

We performed Mendelian randomization for six traits associated with several variants in the phenome-wide association study. We constructed genetic instruments using summary statistics from GWAS performed on UK Biobank (White British) and assessed the association of these instruments with AS. Using the IVW method, genetically elevated levels of apolipoprotein B, lipoprotein(a), LDL-C, and BMI were associated with increased odds of AS [OR per g/L of apolipoprotein B, 3.85; 95% CI, 2.90–5.12; $P = 2.1 \times 10^{-20}$; OR per natural logarithm of lipoprotein(a), 1.20; 95% CI, 1.17–1.23; $P = 3.0 \times 10^{-53}$; OR per mmol/L of LDL-C, 1.57; 95% CI, 1.44–1.71; $P = 1.5 \times 10^{-25}$; OR per kg/m^2 of BMI, 1.07; 95% CI, 1.05, 1.09; $P = 1.9 \times 10^{-12}$, respectively], with no evidence of pleiotropy (all Egger intercepts $P > 0.05$) (see Supplementary material online, *Table S11*). The associations between AS and higher genetically predicted levels of both C-reactive protein and alkaline phosphatase were pleiotropic (Egger intercept P-values of 0.024 and 0.003), and MR results were not consistent across methods limiting causal inference (see Supplementary material online, *Table S11*).

Discussion

In this genome-wide meta-analysis combining summary statistics from 10 cohorts of European ancestry, we confirmed association for seven previously identified loci and identified 10 loci not previously reported to be genome-wide significant for AS. For newly discovered variants, we observed modest changes in AS odds (9%–55% increased odds per risk allele). In independent replication, 8 of the 10 previously unreported variants were associated with AS, with the exception of *BET1L* rs73386631 and *HMGB1* rs181753401 which had low allele frequencies. Our work brings the total number of loci robustly associated with AS to 15. Several of these variants were associated with AS in other ancestries, and a risk score composed of all identified variants was a predictor of AS independent of other risk factors. Further analysis using region- and gene-based methods identified additional gene regions associated with AS, including *LDLR* and *NOTCH4*, as well as

Downloaded from https://academic.oup.com/eurheartj/article/44/21/1927/7112365 by guest on 15 October 2025

differential expression of co-regulated groups of genes. Finally, Mendelian randomization supported a causal contribution of lipoprotein(a), apolipoprotein B, LDL-C, and BMI to AS.

Our findings support four key etiological mechanisms for AS: calcification, lipid metabolism, adiposity, and inflammation (*Structured Graphical Abstract*). The variants at the *PRRX1*, *ACTR2*, *LPA*, *NLRP6*, and *FADS1/2* loci were also associated with aortic valve calcification assessed by computed tomography. We observed associations of the index variants at the *CELSR2–SORT1*, *PRRX1*, *TEX41*, and *FADS1/2* loci with heel bone mineral density, suggesting systemic effects on calcification. The paired-related homeobox protein 1, the product of the locus *PRRX1*, is a transcription factor required for osteoblast differentiation by TNF-α[41] and is an inducer of the epithelial–mesenchymal transition,[42] a key process in early calcification. Interestingly, zinc finger e-box binding homeobox 2, coded by *ZEB2*, is a repressor of this process.[43] Predicted *ZEB2* expression in the aorta was the most significant differentially expressed gene. Consistent with a potential role in calcification, earlier work demonstrated that *TEX41* variants may be associated with AS through long-range chromatin interactions with the *ZEB2* promoter region, including in the aorta[9] implicating *TEX41*, *ZEB2*, and *PRRX1* in inducing early calcification and subsequent valve stenosis.

Located in the 3' untranslated region of *CELSR2*, rs12740374 affects expression of *SORT1*,[44] which has been reported to decrease hepatic excretion of apolipoprotein B and increase catabolism of LDL-C.[45] Studies have reported the association of lower LDL-C[46,47] and lower odds of CAD,[48,49] and in a study involving two cohorts participating in the current study, a variant in perfect LD with rs12740374 was associated with AS after Bonferroni correction ($P = 3.4 \times 10^{-4}$).[9] The present analysis identified genome-wide significance in the association of the *CELSR2–SORT1* variant with AS. Consistent with its effects on LDL-C and apolipoprotein B, the minor allele conferred a reduction in the odds of AS. Additional evidence that lipid metabolism is a causal mechanism for AS was provided by Mendelian randomization, which confirmed the role of LDL-C[12] and identified a contribution of apolipoprotein B, which extends this association to all apolipoprotein B-containing lipoprotein particles. The findings were consistent with work demonstrating an association of a non-high-density lipoprotein cholesterol genetic risk score with AS.[9] In addition to these lipoprotein-mediated effects, sortilin, coded by *SORT1*, is a regulator of vascular calcification and is associated with increased aortic calcification in mice.[50] Furthermore, sortilin expression is associated with increased expression of *ALPL*,[50] a known AS locus replicated in this study.

Both overall and abdominal obesity have been previously associated with AS,[51] and a Mendelian randomization study provided support for the causality of BMI.[14] Our Mendelian randomization analyses replicate and extend these findings by observing genetic correlations between AS and multiple measures of adiposity, including BMI, waist and hip circumferences, and obesity. However, only one of our genome-wide significant variants, *ARHGEF26* rs6794263, which was in high LD with the missense variant rs13096373 (p.Phe203Ser), was associated with BMI and hip circumference in our phenome-wide analysis. Another missense variant rs12493885 (p.Val29Leu) in *ARHGEF26* was previously associated with CAD,[52] mediated by a gain of function for ARHGEF26 that may lead to increased transendothelial migration, greater adhesion of leukocytes, and proliferation of vascular smooth muscle cells.[52] However, this variant was independent of *ARHGEF26* rs6794263 ($r^2 = 0.018$)[39] and was not associated with AS in our meta-analysis (OR per T allele, 1.02; 95% CI, 0.98–1.06; $P = 0.44$ for rs1713812, which is in perfect LD with rs12493885[39]). Conversely, the allele of rs6794263 that confers a lower odds of AS has also been associated with lower odds of CAD, but not at genome-wide significance (OR per C allele, 0.96; 95% CI, 0.92–1.00; $P = 0.037$).[53] The presence of two independent, missense mutations in *ARHGEF26* with discordant effects on AS and CAD suggests pleiotropic effects of this locus.

The accumulation of inflammatory cells in the aortic valve is associated with remodeling and fibrosis,[54] highlighting the role for inflammation in disease progression. We confirmed that the *FADS1/2* locus was associated with AS.[11] Both *FADS1* and *FADS2* encode key enzymes in the conversion of dietary *n*-6 fatty acids to arachidonic acid, a precursor of pro-inflammatory leukotrienes and prostaglandins.[55] In addition, we also identified a novel variant at the locus for NLRP6, which assembles an inflammasome that plays a role in immunity to bacterial infection as well as proinflammatory responses to other stimuli, including fatty acids.[56,57] We replicated that the association previously reported between an *IL6* variant and AS[10] and the *IL6* variant rs2069832 ($r^2 = 0.95$ with our index variant) colocalizes with *IL6-AS1* expression.[10] Notably, the risk-increasing alleles are associated with increased expression of both *IL6* and *IL6-AS1* in fibroblasts in GTEx.[23] Our analyses therefore indicate a pro-inflammatory association with AS in the *IL6* region but also higher predicted *IL6R* expression in the blood cells of AS cases. Thus, several orthogonal signals support the association of pro-inflammatory pathways with AS.

The present study represents the largest genome-wide meta-analysis of AS, a common condition with no available medical treatment, and included 653 867 participants (13 765 cases) in which we applied variant-, gene-, and gene set-based analyses to identify additional risk loci and disease mechanisms. Our work highlights novel mechanisms and pathways which may have important clinical and future research implications. First, our analyses point to several possible therapeutic interventions using both lifestyle and novel drugs. Our findings of lipoproteins and adiposity as key drivers of aortic stenosis suggest that maintenance of healthy lifestyle and adherence to lipid-lowering recommendations may reduce the incidence of aortic stenosis. Furthermore, lipoprotein(a)-lowering therapies currently in development[58,59] may provide a new avenue for prevention of disease progression. Whether novel hypoglycemics, which also lead to marked weight loss and metabolic improvements,[60] and anti-inflammatory agents (e.g. that inhibit IL-6 signaling[61]) could reduce AS requires randomized trials. Finally, our observation of a robust association between a genetic risk score and AS may allow identification of patients at risk for rapid disease progression in clinical practice, who may require echocardiographic follow-up, as well as permit targeted enrolment of such at-risk patients in future randomized trials.

Despite the strengths of the study, there are several limitations. Since our discovery cohorts were of European ancestry, the transferability of our results to other ancestries may be limited.[62–64] Although we attempted cross-ancestry replication of genome-wide significant variants, the observed limited reproducibility may be due to the low numbers of African and Latin American participants. Future studies should focus on non-Europeans. Our analyses also made exclusive use of bioinformatic methods to identify genetic loci. These results require confirmation using complementary approaches. Lastly, cases were selected using various criteria (international classification of diseases (ICD) codes, surgical AVR, etc.) and not solely by echocardiography. Therefore we could not entirely exclude bicuspid aortic valves; however, these likely represent a small proportion of the cases.

In conclusion, our results identify novel genetic contributors to AS and identify specific contributions to disease etiology that are characterized by the effects of calcification, altered lipid metabolism, adiposity, and inflammation. An AS genetic risk score was an independent predictor of both clinical and subclinical disease, providing additional discrimination when added to clinical risk factors. Established and novel

Downloaded from https://academic.oup.com/eurheartj/article/44/21/1927/7112365 by guest on 15 October 2025

genetic loci warrant investigation as potential therapeutic targets to prevent the initiation of aortic calcification and progression to stenosis.

Author contributions

Concept and design: Chen, Engert, and Thanassoulis. *Acquisition, analysis, or interpretation of data:* Chen, Dina, Small, Shaffer, Levinson, Helgadóttir, Capoulade, Munter, Martinsson, Cairns, Trudsø, Hoekstra, Burr, Marsh, Dufresne, Messika-Zeitoun, Le Scouarnec, Ghouse, Olesen, Christensen, Mikkelsen, Jacobsen, Dowsett, Pedersen, Erikstrup, Ostrowski, Budoff, Post, Rotter, Bundgaard, Le Tourneau, Smith, Hólm, Söderberg, Schott, Engert, and Thanassoulis. *Drafted the manuscript:* Chen. *Critical revision of the manuscript for important intellectual content:* Helgadóttir, Capoulade, Martinsson, Cairns, Trudsø, Hoekstra, Marsh, Dufresne, Hólm, Christensen, Mikkelsen, Erikstrup, Ostrowski, Post, Rotter, Bundgaard, Johansson, Ljungberg, Näslund, Smith, Söderberg, Schott, Clarke, Engert, and Thanassoulis. *Obtained funding:* Gudnason, Näslund, Post, Rotter, Lathrop, Le Tourneau, Messika-Zeitoun, Smith, Söderberg, Schott, Engert, and Thanassoulis. *Administrative, technical, or material support:* Ranatunga, Whitmer, Bonnefond, Lathrop, Ljungberg, Näslund, Le Tourneau, Smith, Söderberg, Schott, Engert, and Thanassoulis. *Supervision:* Damrauer, Budoff, Gudnason, Rotter, Johansson, Smith, Rader, Clarke, Engert, and Thanassoulis.

Acknowledgements

We would like to acknowledge the valuable contributions of the investigators, staff, and participants from all cohorts. Analyses conducted using the UK Biobank Resource were performed under Application Numbers 11537, 41025, and 24281.

Supplementary data

Supplementary data is available at *European Heart Journal* online.

Data availability

The summary level statistics from the meta-GWAS that support the findings of this study are available online @ Zenodo.org (https://doi.org/10.5281/zenodo.7505361).

Conflict of interest

Scott M. Damrauer receives research support (to the University of Pennsylvania) from RenalytixAI and personal fees from Caico Ibs, both outside the scope of the present work. SMD is also named as a co-inventor on a government-owned US Patent application related to the use of genetic risk prediction for venous thromboembolic disease filed by the US Department of Veterans Affairs in accordance with Federal regulatory requirements. SMD is named as a co-inventor on a Government-owned US Patent application related to the use of PDE3B inhibition for preventing cardiovascular disease filed by the US Department of Veterans Affairs in accordance with Federal regulatory requirements. Stefan Söderberg has received speaker honoraria and consulting fees from Actelion Ltd. George Thanassoulis has received consulting fees from Ionis Pharmaceuticals and has participated in advisory boards for Amgen, Sanofi, Novartis, HLS Therapeutics and Silence. Morten Salling Olesen has received 5.000.000 dkrfra Sundhedsdonationer.Journalnr. 2022-0243. David O. Arnar has received travel support from Pfizer to attend the ESC 2022 Scientific Meeting in Barcelona and has stock options in Sidekick Health Digital Therapeutics. Henning Bundgaard has received lecture fees from Amgen, MSD, Sanofi-Avensis, BMS and grants from NordForsk, Innovation Fond, Denmark, The Capital Regions Research Foundation. Alex Hoerby Christensen—Novo Nordisk Foundation NNF20OC0065799. Romaine Capoulade has received an Honorarium for one lecture from Novartis. Robert Clarke has received support from BAYER (China Kadoorie Biobank). Unnur Thorsteinsdottir's research is funded by deCODE genetics/Amgen. Daniel F. Gudbjartsson receives funds from deCODE Genetics/Amgen. Until 1 June 2022, Gudmundur Thorgeirsson was a part time employee of deCode Genetics that is owned by Amgen. Hilma Holm is an employee of deCODE genetics/Amgen Inc. Anna Helgadottir is an employee of deCODE genetics/Amgen Inc.

Funding

Thomas W. Marsh was supported by a research award from the Fonds de Recherche du Québec—Santé. Scott M. Damrauer is supported by the US Department of Veteran Affairs (IK2-CX00180). This work does not represent the views of the US Government or the US Department of Veterans Affairs. Stefan Söderberg was supported by Hjärt-Lungfonden (grant numbers 20140799, 20120631, and 20100635), the County Council of Västerbotten (ALF, VLL-548791), Umeå University, and the Heart Foundation of Northern Sweden. J. Gustav Smith was supported by grants from the Swedish Heart-Lung Foundation (2016-0134, 2016-0315, and 2019-0526), the Swedish Research Council (2017-02554), the European Research Council (ERC-STG-2015-679242), the Crafoord Foundation, Skåne University Hospital, the Scania County, governmental funding of clinical research within the Swedish National Health Service, a generous donation from the Knut and Alice Wallenberg foundation to the Wallenberg Center for Molecular Medicine in Lund, and funding from the Swedish Research Council (Linnaeus grant Dnr 349-2006-237, Strategic Research Area Exodiab Dnr 2009-1039) and the Swedish Foundation for Strategic Research (Dnr IRC15-0067) to the Lund University Diabetes Center. George Thanassoulis is a Chercheur Boursier Clinicien—Senior from the Fonds de Recherche du Québec—Santé. George Thanassoulis and James C. Engert hold an operating grant from the Canadian Institutes for Health Research, and George Thanassoulis holds operating grants from the Heart and Stroke Foundation of Canada and the National Institutes of Health (R01 HL128550). This study was funded by grants from the Fonds de Recherche du Québec—Santé, Canadian Institutes for Health Research, Heart and Stroke Foundation of Canada, and National Institutes of Health (R01 HL128550). The Kaiser Permanente Research Program on Genes, Environment, and Health was funded by the Ellison Medical Foundation, Robert Wood Johnson Foundation, Wayne and Gladys Valley Foundation, Kaiser Permanente Northern California, and the Kaiser Permanente National and Regional Community Benefit Programs. The GERA cohort was funded by a grant from the National Institutes of Health. The University of Oxford MRC Population Health Research Unit is funded through a strategic partnership between the Medical Research Council and the University of Oxford. Staff were also supported by the British Heart Foundation and British Heart Foundation Oxford Centre for Research Excellence (RE/18/3/34214). A full list of participating MESA investigators and institutions can be found at http://www.mesa-nhlbi.org. Multi-Ethnic Study of Atherosclerosis and the MESA SHARe project are conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators. Support for MESA is provided by contracts 75N92020D00001, HHSN268201500003I, N01-HC-95159, 75N92020D00005, N01-HC-95160, 75N92020D00002, N01-HC-95161, 75N92020D00003, N01-HC-95162, 75N92020D00006, N01-HC-95163, 75N92020D00004, N01-HC-95164, 75N92020D00007, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168, N01-HC-95169, UL1-TR-000040, UL1-TR-001079, UL1-TR-001420. Funding for SHARe genotyping was provided by NHLBI Contract N02-HL-64278. Genotyping was performed at Affymetrix (Santa Clara, California, USA) and the Broad Institute of Harvard and MIT (Boston,

Downloaded from https://academic.oup.com/eurheartj/article/44/21/1927/7112365 by guest on 15 October 2025

Massachusetts, USA) using the Affymetrix Genome-Wide Human SNP Array 6.0. Also supported in part by the National Center for Advancing Translational Sciences, CTSI grant UL1TR001881, and the National Institute of Diabetes and Digestive and Kidney Disease Diabetes Research Center (DRC) grant DK063491 to the Southern California Diabetes Endocrinology Research Center. This research was supported by R01 HL071739. Infrastructure for the CHARGE consortium is supported in part by the NHLBI grant R01HL105756 and also supported in part by National Institute of Health contract R01HL146860. Vanderbilt University Medical Center's BioVU projects are supported by numerous sources: institutional funding, private agencies, and federal grants. These include the National Institutes of Health Grant S10RR025141; Clinical and Translational Science Award grants UL1TR002243, UL1TR000445, and UL1RR024975. Genomic data are also supported by investigator-led projects that include U01HG004798, R01NS032830, RC2GM092618, P50GM115305, U01HG006378, U19HL065962, and R01HD07471. This work was supported by an ANR & FRM grant [13-BSV6-0011 and DCV20070409278] to Jean-Jacques Schott, by a Fédération Française de Cardiologie, a Fondation Coeur et Recherche and an Inserm Translational Research grant to Thierry Le Tourneau. Romain Capoulade is supported by a 'Connect Talent' research chair from Région Pays de la Loire. Cohort PREGI set up was supported by the French Regional Council of Pays-de-la-Loire (VaCaRMe program). We thank Marie Marrec and Guénola Coste for their contribution to clinical data collection and the Genomics and Bioinformatics Core Facility of Nantes (GenoBiRD, Biogenouest) for its technical support. The D.E.S.I.R. study (providing a set of controls) has been supported by INSERM contracts with CNAMTS, Lilly, Novartis Pharma, and Sanofi-Aventis; by INSERM (Réseaux en Santé Publique, Interactions entre les déterminants de la santé), Cohortes Santé TGIR, the Association Diabète Risque Vasculaire, the Fédération Française de Cardiologie, La Fondation de France, ALFEDIAM, Société francophone du diabète, ONIVINS, Abbott, Ardix Medical, Bayer Diagnostics, Becton Dickinson, Cardionics, Merck Santé, Novo Nordisk, Pierre Fabre, Roche, and Topcon. The D.E.S.I.R. Study Group. INSERM U1018: B. Balkau, P. Ducimetière, E. Eschwège; INSERM U367: F. Alhenc-Gelas; CHU D'Angers: A. Girault; Bichat Hospital: F. Fumeron, M. Marre, R Roussel; CHU de Rennes: F. Bonnet; CNRS UMR8090, Lille: A. Bonnefond, P. Froguel; Univ Paris Descartes, UMR1153, Paris: F.Rancière Centres d'Examens de Santé: Alençon, Angers, Blois, Caen, Chateauroux, Chartres, Cholet, Le Mans, Orléans, Tours; Institute de Recherche Médecine Générale: J. Cogneau; General practitioners of the Region; Institute inter-Regional pour la Santé: C. Born, E. Caces, M. Cailleau, O Lantieri, J.G. Moreau, F. Rakotozafy, J. Tichet, S. Vol. We would like to specially thank the team of the Centre d'Investigation Clinique (Xavier Duval), the Centre de Ressources Biologiques (Sarah Tubiana), Christophe Aucan from the Assistance Publique—Hôpitaux de Paris, Département de la Recherche Clinique et du Développement (DRCD) and Estelle Marcault from the Unité de Recherche Clinique Paris Nord for their help and support during all these years. The COFRASA (clinicalTrial.gov number NCT 00338676) and GENERAC (clinicalTrial.gov number NCT00647088) studies are supported by grants from the Assistance Publique—Hôpitaux de Paris (PHRC National 2005 and 2010, and PHRC regional 2007). The Copenhagen Hospital Biobank and The Danish Blood Donor Study were supported by the Novo Nordisk Foundation (grant numbers NNF14CC0001 and NNF17OC0027594). This work was further supported by The Innovation Fund Denmark (PM Heart) Nord-Forsk (project no. 90580) and the Capital Regions Research Council (No. A5920). None of the funding sources were involved in the collection, analysis, and interpretation of data, nor the writing of this paper or the decision to submit the paper for publication.

References

1. Andell P, Li X, Martinsson A, Andersson C, Stagmo M, Zoller B, *et al.* Epidemiology of valvular heart disease in a Swedish nationwide hospital-based register study. *Heart* 2017; **103**:1696–1703. https://doi.org/10.1136/heartjnl-2016-310894
2. Danielsen R, Aspelund T, Harris TB, Gudnason V. The prevalence of aortic stenosis in the elderly in Iceland and predictions for the coming decades: the AGES-Reykjavik study. *Int J Cardiol* 2014;**176**:916–922. https://doi.org/10.1016/j.ijcard.2014.08.053
3. Vahanian A, Beyersdorf F, Praz F, Milojevic M, Baldus S, Bauersachs J, *et al.* 2021 ESC/EACTS guidelines for the management of valvular heart disease. *Eur Heart J* 2022;**43**: 561–632. https://doi.org/10.1093/eurheartj/ehab395
4. Members WC, Otto CM, Nishimura RA, Bonow RO, Carabello BA, Erwin JP III, *et al.* 2020 ACC/AHA guideline for the management of patients with valvular heart disease: a report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. *J Am Coll Cardiol* 2021;**77**:e25–e197. https://doi.org/10.1016/j.jacc.2020.11.018
5. Martinsson A, Li X, Zoller B, Andell P, Andersson C, Sundquist K, *et al.* Familial aggregation of aortic valvular stenosis: a nationwide study of sibling risk. *Circ Cardiovasc Genet* 2017;**10**:e001742. https://doi.org/10.1161/CIRCGENETICS.117.001742
6. Probst V, Le Scouarnec S, Legendre A, Jousseaume V, Jaafar P, Nguyen JM, *et al.* Familial aggregation of calcific aortic valve stenosis in the western part of France. *Circulation* 2006;**113**:856–860. https://doi.org/10.1161/CIRCULATIONAHA.105.569467
7. Thanassoulis G, Campbell CY, Owens DS, Smith JG, Smith AV, Peloso GM, *et al.* Genetic associations with valvular calcification and aortic stenosis. *N Engl J Med* 2013; **368**:503–512. https://doi.org/10.1056/NEJMoa1109034
8. Theriault S, Gaudreault N, Lamontagne M, Rosa M, Boulanger MC, Messika-Zeitoun D, *et al.* A transcriptome-wide association study identifies PALMD as a susceptibility gene for calcific aortic valve stenosis. *Nat Commun* 2018;**9**:988. https://doi.org/10.1038/s41467-018-03260-6
9. Helgadottir A, Thorleifsson G, Gretarsdottir S, Stefansson OA, Tragante V, Thorolfsdottir RB, *et al.* Genome-wide analysis yields new loci associating with aortic valve stenosis. *Nat Commun* 2018;**9**:987. https://doi.org/10.1038/s41467-018-03252-6
10. Theriault S, Dina C, Messika-Zeitoun D, Le Scouarnec S, Capoulade R, Gaudreault N, *et al.* Genetic association analyses highlight IL6, ALPL, and NAV1 as 3 new susceptibility genes underlying calcific aortic valve stenosis. *Circ Genom Precis Med* 2019;**12**:e002617. https://doi.org/10.1161/CIRCGEN.119.002617
11. Chen HY, Cairns BJ, Small AM, Burr HA, Ambikkumar A, Martinsson A, *et al.* Association of FADS1/2 locus variants and polyunsaturated fatty acids with aortic stenosis. *JAMA Cardiol* 2020;**5**:694–702. https://doi.org/10.1001/jamacardio.2020.0246
12. Smith JG, Luk K, Schulz CA, Engert JC, Do R, Hindy G, *et al.* Association of low-density lipoprotein cholesterol-related genetic variants with aortic valve calcium and incident aortic stenosis. *JAMA* 2014;**312**:1764–1771. https://doi.org/10.1001/jama.2014.13959
13. Arsenault BJ, Boekholdt SM, Dube MP, Rheaume E, Wareham NJ, Khaw KT, *et al.* Lipoprotein(a) levels, genotype, and incident aortic valve stenosis: a prospective Mendelian randomization study and replication in a case-control cohort. *Circ Cardiovasc Genet* 2014;**7**:304–310. https://doi.org/10.1161/CIRCGENETICS.113.000400
14. Kaltoft M, Langsted A, Nordestgaard BG. Obesity as a causal risk factor for aortic valve stenosis. *J Am Coll Cardiol* 2020;**75**:163–176. https://doi.org/10.1016/j.jacc.2019.10.050
15. Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM; 1000 Genomes Project Consortium, *et al.* A global reference for human genetic variation. *Nature* 2015;**526**: 68–74. https://doi.org/10.1038/nature15393
16. Gonzalez JN, Zweig AS, Speir ML, Schmelter D, Rosenbloom KR, Raney BJ, *et al.* The UCSC genome browser database: 2021 update. *Nucleic Acids Res* 2021;**49**: D1046–D1057. https://doi.org/10.1093/nar/gkaa1070
17. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 2010;**38**:e164. https://doi.org/10.1093/nar/gkq603
18. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res* 2019;**47**: D886–D894. https://doi.org/10.1093/nar/gky1016
19. Quang D, Chen Y, Xie X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* 2015;**31**:761–763. https://doi.org/10.1093/bioinformatics/btu703
20. Huang YF, Gulko B, Siepel A. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nat Genet* 2017;**49**:618–624. https://doi.org/10.1038/ng.3810
21. Ionita-Laza I, McCallum K, Xu B, Buxbaum JD. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat Genet* 2016;**48**:214–220. https://doi.org/10.1038/ng.3477
22. Shihab HA, Gough J, Mort M, Cooper DN, Day IN, Gaunt TR. Ranking non-synonymous single nucleotide polymorphisms based on disease concepts. *Hum Genomics* 2014;**8**:11. https://doi.org/10.1186/1479-7364-8-11
23. GTEx Consortium. The genotype-tissue expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* 2015;**348**:648–660. https://doi.org/10.1126/science.1262110
24. de Leeuw CA, Mooij JM, Heskes T, Posthuma D. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput Biol* 2015;**11**:e1004219. https://doi.org/10.1371/journal.pcbi.1004219
25. Barbeira AN, Dickinson SP, Bonazzola R, Zheng J, Wheeler HE, Torres JM, *et al.* Exploring the phenotypic consequences of tissue specific gene expression variation

Downloaded from https://academic.oup.com/eurheartj/article/44/21/1927/7112365 by guest on 15 October 2025

inferred from GWAS summary statistics. *Nat Commun* 2018;**9**:1825. https://doi.org/10.1038/s41467-018-03621-1

26. Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, *et al.* Bayesian Test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet* 2014;**10**:e1004383. https://doi.org/10.1371/journal.pgen.1004383
27. Zhu S, Qian T, Hoshida Y, Shen Y, Yu J, Hao K. GIGSEA: genotype imputed gene set enrichment analysis using GWAS summary level data. *Bioinformatics* 2019;**35**:160–163. https://doi.org/10.1093/bioinformatics/bty529
28. Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000;**28**:27–30. https://doi.org/10.1093/nar/28.1.27
29. The Gene Ontology Consortium. The gene ontology resource: 20 years and still GOing strong. *Nucleic Acids Res* 2019;**47**:D330–D338. https://doi.org/10.1093/nar/gky1055
30. Lizio M, Harshbarger J, Shimoji H, Severin J, Kasukawa T, Sahin S, *et al.* Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biol* 2015;**16**:22. https://doi.org/10.1186/s13059-014-0560-6
31. Kozomara A, Birgaoanu M, Griffiths-Jones S. miRBase: from microRNA sequences to function. *Nucleic Acids Res* 2019;**47**:D155–D162. https://doi.org/10.1093/nar/gky1141
32. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 2015;**4**:7. https://doi.org/10.1186/s13742-015-0047-8
33. Prive F, Arbel J, Vilhjalmsson BJ. LDpred2: better, faster, stronger. *Bioinformatics* 2020;**36**:5424–5431. https://doi.org/10.1093/bioinformatics/btaa1029
34. Yavorska OO, Burgess S. Mendelianrandomization: an R package for performing Mendelian randomization analyses using summarized data. *Int J Epidemiol* 2017;**46**:1734–1739. https://doi.org/10.1093/ije/dyx034
35. Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J; Schizophrenia Working Group of the Psychiatric Genomics Consortium, *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* 2015;**47**:291–295. https://doi.org/10.1038/ng.3211
36. Zheng J, Erzurumluoglu AM, Elsworth BL, Kemp JP, Howe L, Haycock PC, *et al.* LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics* 2017;**33**:272–279. https://doi.org/10.1093/bioinformatics/btw613
37. Yang J, Ferreira T, Morris AP, Medland SE; Genetic Investigation of ANthropometric Traits (GIANT) Consortium; DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium, *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat Genet* 2012;**44**:369–375, S361–363. https://doi.org/10.1038/ng.2213
38. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* 2011;**88**:76–82. https://doi.org/10.1016/j.ajhg.2010.11.011
39. Machiela MJ, Chanock SJ. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics* 2015;**31**:3555–3557. https://doi.org/10.1093/bioinformatics/btv402
40. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* 2014;**46**:310–315. https://doi.org/10.1038/ng.2892
41. Lu X, Beck GR Jr, Gilbert LC, Camalier CE, Bateman NW, Hood BL, *et al.* Identification of the homeobox protein Prx1 (MHox, prrx-1) as a regulator of osterix expression and mediator of tumor necrosis factor alpha action in osteoblast differentiation. *J Bone Miner Res* 2011;**26**:209–219. https://doi.org/10.1002/jbmr.203
42. Ocana OH, Corcoles R, Fabra A, Moreno-Bueno G, Acloque H, Vega S, *et al.* Metastatic colonization requires the repression of the epithelial-mesenchymal transition inducer Prrx1. *Cancer Cell* 2012;**22**:709–724. https://doi.org/10.1016/j.ccr.2012.10.012
43. Vandewalle C, Comijn J, De Craene B, Vermassen P, Bruyneel E, Andersen H, *et al.* SIP1/ZEB2 Induces EMT by repressing genes of different epithelial cell-cell junctions. *Nucleic Acids Res* 2005;**33**:6566–6578. https://doi.org/10.1093/nar/gki965
44. Musunuru K, Strong A, Frank-Kamenetsky M, Lee NE, Ahfeldt T, Sachs KV, *et al.* From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature* 2010;**466**:714–719. https://doi.org/10.1038/nature09266
45. Strong A, Ding Q, Edmondson AC, Millar JS, Sachs KV, Li X, *et al.* Hepatic sortilin regulates both apolipoprotein B secretion and LDL catabolism. *J Clin Invest* 2012;**122**:2807–2816. https://doi.org/10.1172/JCI63563
46. Hoffmann TJ, Theusch E, Haldar T, Ranatunga DK, Jorgenson E, Medina MW, *et al.* A large electronic-health-record-based genome-wide study of serum lipids. *Nat Genet* 2018;**50**:401–413. https://doi.org/10.1038/s41588-018-0064-5
47. Kathiresan S, Willer CJ, Peloso GM, Demissie S, Musunuru K, Schadt EE, *et al.* Common variants at 30 loci contribute to polygenic dyslipidemia. *Nat Genet* 2009;**41**:56–65. https://doi.org/10.1038/ng.291
48. van der Harst P, Verweij N. Identification of 64 novel genetic loci provides an expanded view on the genetic architecture of coronary artery disease. *Circ Res* 2018;**122**:433–443. https://doi.org/10.1161/CIRCRESAHA.117.312086
49. Nikpay M, Goel A, Won HH, Hall LM, Willenborg C, Kanoni S, *et al.* A comprehensive 1,000 genomes-based genome-wide association meta-analysis of coronary artery disease. *Nat Genet* 2015;**47**:1121–1130. https://doi.org/10.1038/ng.3396
50. Goettsch C, Hutcheson JD, Aikawa M, Iwata H, Pham T, Nykjaer A, *et al.* Sortilin mediates vascular calcification via its recruitment into extracellular vesicles. *J Clin Invest* 2016;**126**:1323–1336. https://doi.org/10.1172/JCI80851
51. Larsson SC, Wolk A, Hakansson N, Back M. Overall and abdominal obesity and incident aortic valve stenosis: two prospective cohort studies. *Eur Heart J* 2017;**38**:2192–2197. https://doi.org/10.1093/eurheartj/ehx140
52. Klarin D, Zhu QM, Emdin CA, Chaffin M, Horner S, McMillan BJ, *et al.* Genetic analysis in UK Biobank links insulin resistance and transendothelial migration pathways to coronary artery disease. *Nat Genet* 2017;**49**:1392–1397. https://doi.org/10.1038/ng.3914
53. Schunkert H, Konig IR, Kathiresan S, Reilly MP, Assimes TL, Holm H, *et al.* Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nat Genet* 2011;**43**:333–338. https://doi.org/10.1038/ng.784
54. Cote N, Mahmut A, Bosse Y, Couture C, Page S, Trahan S, *et al.* Inflammation is associated with the remodeling of calcific aortic valve disease. *Inflammation* 2013;**36**:573–581. https://doi.org/10.1007/s10753-012-9579-6
55. Schmitz G, Ecker J. The opposing effects of n-3 and n-6 fatty acids. *Prog Lipid Res* 2008;**47**:147–155. https://doi.org/10.1016/j.plipres.2007.12.004
56. Zheng D, Kern L, Elinav E. The NLRP6 inflammasome. *Immunology* 2021;**162**:281–289. https://doi.org/10.1111/imm.13293
57. Lee HJ, Yeon JE, Ko EJ, Yoon EL, Suh SJ, Kang K, *et al.* Peroxisome proliferator-activated receptor-delta agonist ameliorated inflammasome activation in nonalcoholic fatty liver disease. *World J Gastroenterol* 2015;**21**:12787–12799. https://doi.org/10.3748/wjg.v21.i45.12787
58. Koren MJ, Moriarty PM, Baum SJ, Neutel J, Hernandez-Illas M, Weintraub HS, *et al.* Preclinical development and phase 1 trial of a novel siRNA targeting lipoprotein(a). *Nat Med* 2022;**28**:96–103. https://doi.org/10.1038/s41591-021-01634-w
59. Tsimikas S, Karwatowska-Prokopczuk E, Gouni-Berthold I, Tardif JC, Baum SJ, Steinhagen-Thiessen E, *et al.* Lipoprotein(a) reduction in persons with cardiovascular disease. *N Engl J Med* 2020;**382**:244–255. https://doi.org/10.1056/NEJMoa1905239
60. Malik IO, Petersen MC, Klein S. Glucagon-like peptide-1, glucose-dependent insulinotropic polypeptide, and glucagon receptor poly-agonists: a new era in obesity pharmacotherapy. *Obesity (Silver Spring)* 2022;**30**:1718–1721. https://doi.org/10.1002/oby.23521
61. Hafiane A, Daskalopoulou SS. Targeting the residual cardiovascular risk by specific anti-inflammatory interventions as a therapeutic strategy in atherosclerosis. *Pharmacol Res* 2022;**178**:106157. https://doi.org/10.1016/j.phrs.2022.106157
62. Han Y, Dorajoo R, Chang X, Wang L, Khor CC, Sim X, *et al.* Genome-wide association study identifies a missense variant at APOA5 for coronary artery disease in multi-ethnic cohorts from Southeast Asia. *Sci Rep* 2017;**7**:17921. https://doi.org/10.1038/s41598-017-18214-z
63. Adeyemo A, Bentley AR, Meilleur KG, Doumatey AP, Chen G, Zhou J, *et al.* Transferability and fine mapping of genome-wide associated loci for lipids in African Americans. *BMC Med Genet* 2012;**13**:88. https://doi.org/10.1186/1471-2350-13-88
64. Liu CT, Raghavan S, Maruthur N, Kabagambe EK, Hong J, Ng MC, *et al.* Trans-ethnic meta-analysis and functional annotation illuminates the genetic architecture of fasting glucose and insulin. *Am J Hum Genet* 2016;**99**:56–75. https://doi.org/10.1016/j.ajhg.2016.05.006

Downloaded from https://academic.oup.com/eurheartj/article/44/21/1927/7112365 by guest on 15 October 2025

REVIEW

Plasticity of the *Leishmania* genome leading to gene copy number variations and drug resistance [version 1; referees: 5 approved]

Marie-Claude N. Laffitte*, Philippe Leprohon*, Barbara Papadopoulou, Marc Ouellette

Centre de Recherche en Infectiologie du Centre de Recherche du CHU Québec, and Département de Microbiologie, Infectiologie et Immunologie, Faculté de Médecine, Université Laval, Québec, Québec, Canada

* Equal contributors

v1 **First published:** 20 Sep 2016, **5**(F1000 Faculty Rev):2350 (doi: 10.12688/f1000research.9218.1)

Latest published: 20 Sep 2016, **5**(F1000 Faculty Rev):2350 (doi: 10.12688/f1000research.9218.1)

Abstract

Leishmania has a plastic genome, and drug pressure can select for gene copy number variation (CNV). CNVs can apply either to whole chromosomes, leading to aneuploidy, or to specific genomic regions. For the latter, the amplification of chromosomal regions occurs at the level of homologous direct or inverted repeated sequences leading to extrachromosomal circular or linear amplified DNAs. This ability of *Leishmania* to respond to drug pressure by CNVs has led to the development of genomic screens such as Cos-Seq, which has the potential of expediting the discovery of drug targets for novel promising drug candidates.

Open Peer Review

Referee Status: ✓ ✓ ✓ ✓ ✓

	Invited Referees			
	1	2	3	4
version 1 published 20 Sep 2016	✓	✓	✓	✓
	5			
	✓			

F1000 Faculty Reviews are commissioned from members of the prestigious F1000 Faculty. In order to make these reviews as comprehensive and accessible as possible, peer review takes place before publication; the referees are listed below, but their reports are not formally published.

1 **Francisco Gamarro**, Instituto de Parasitología Lopez-Neyra, IPBLN-CSIC Spain

2 **Joachim Clos**, Bernhard Nocht Institute for Tropical Medicine Germany

3 **Richard McCulloch**, University of Glasgow UK

4 **Yvon Sterkers**, University Regional Hospital of Montpellier France

5 **Gerald Spaeth**, Institut Pasteur, INSERM U1201 France

Discuss this article

Comments (0)

Corresponding authors: Barbara Papadopoulou (barbara.papadopoulou@crchudequebec.ulaval.ca), Marc Ouellette (marc.ouellette@crchudequebec.ulaval.ca)

How to cite this article: Laffitte MCN, Leprohon P, Papadopoulou B and Ouellette M. **Plasticity of the *Leishmania* genome leading to gene copy number variations and drug resistance [version 1; referees: 5 approved]** *F1000Research* 2016, **5**(F1000 Faculty Rev):2350 (doi: 10.12688/f1000research.9218.1)

Copyright: © 2016 Laffitte MCN *et al*. This is an open access article distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Grant information: BP is supported by grants from the Canadian Institutes of Health Research, the Natural Sciences and Engineering Research Council of Canada, and the Ministère du Developpement Économique de l'Innovation et de l'Exportation du Québec. MO is supported by grants from the Canadian Institutes of Health Research and holds a Canada Research Chair on Antimicrobial Resistance.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors declare that they have no competing interests.

First published: 20 Sep 2016, **5**(F1000 Faculty Rev):2350 (doi: 10.12688/f1000research.9218.1)

Introduction

Leishmania are dimorphic parasites living as extracellular promastigotes in the digestive tract of *Phlebotomus* or *Lutzomyia* sandflies and as intracellular amastigotes within phagocytic cells (mainly macrophages) of the vertebrate hosts. *Leishmania* species cause leishmaniasis, the second largest parasite killer; there are 1.3 million new cases annually, 12 million people are affected worldwide, and 350 million people are currently at risk[1]. The *Leishmania* genus includes several species, among which more than 20 are pathogenic to humans[2]. *Leishmania* can be divided into two subgenera: the *Leishmania Leishmania* subgenus responsible for visceral and cutaneous leishmaniasis and the *Leishmania Viannia* subgenus often associated with either cutaneous or muco-cutaneous forms of the disease[2]. Visceral leishmaniasis is mainly caused by *L. donovani* and *L. infantum* and is characterized by fever, hepatosplenomegaly, and pancytopenia[3], making it the most severe and deadly form of the disease when compared with the self-healing but nonetheless debilitating skin lesions of cutaneous leishmaniasis[4,5].

No effective human vaccine is currently available against *Leishmania* (a canine *Leishmania* vaccine was recently registered in Europe[6]), and control measures mainly involve chemotherapy[7–9]. Pentavalent antimony (Sb) has been the standard drug for 70 years and remains the mainstay in many endemic regions, apart from Northern India, where antimonial formulations have been rendered obsolete because of widespread parasite resistance. Other first-line therapies include the polyene antibiotic amphotericin B (AMB) for which a single dose was shown to be 95% effective against visceral leishmaniasis in India[10]. Liposomal AMB has become a standard treatment in many countries[11] but requires administration by the intravenous route. Geographical differences in the response to liposomal AMB were reported, and visceral leishmaniasis cases in India were more responsive than those from East Africa or South America[12]. Clinical AMB resistance is scarce and parasites remained susceptible even after multiple rounds of treatment in the same patient[13]. The alkyl-lysophospholipid analogue miltefosine (MTF) is the only oral drug against *Leishmania*[14,15]. It has been successfully used for the treatment of visceral leishmaniasis since its registration in 2002 in India[16]. However, relapse rates are on the rise in India[17,18] and Nepal[19,20], making MTF resistance a likely problem. The aminoglycoside paromomycin (PMM) is also approved for the treatment of visceral leishmaniasis in India[21,22]. So far, the scarce use of PMM has limited the emergence of resistance, but geographical variations in PMM efficacy against visceral leishmaniasis were noted between East Africa (especially Sudan) and India[23,24]. Lastly, pentamidine (PTD) has been abandoned for the treatment of visceral leishmaniasis because of serious toxicities and is mainly restricted to patients with cutaneous leishmaniasis in South America[25–27].

Despite six decades of use, the mode of action (MOA) of antimonials is not known. It has been shown to lead to the production of reactive oxygen species[28–31], the depletion of trypanothione[32], and apoptosis-like death[33–36], but an exact MOA is still awaited. The same applies for MTF, AMB, PTD, and PMM, with the possible exception of AMB, which kills *Leishmania* by forming pores in ergosterol-containing membranes. New molecules with well-defined drug targets are clearly needed.

Leishmania and its genome

The *Leishmania* genome is around 32 Mb and displays over 8,300 coding genes[37,38]. Within the *Leishmania* genus, gene synteny is conserved for more than 99% of genes between *L. major*, *L. infantum*, and *L. braziliensis*, and only few species-specific genes were found[38]. *Leishmania* species have between 34 and 36 chromosomes ranging in size from 0.3 to 2.8 Mb[37–41]. One unique feature characterizing trypanosomatid parasites lies in their genome architecture, their protein-coding genes being organized as large polycistronic units[42,43]. In the absence of defined RNA polymerase II promoters, transcription of the long polycistronic units occurs in a bidirectional fashion from transcriptional start sites located at strand switch regions[43,44]. Processing into individual messenger RNAs (mRNAs) occurs by the addition through *trans*-splicing of a spliced leader RNA (39 nt) to the 5′ ends of each mRNA, coupled to 3′ end polyadenylation[45,46]. Because of its lack of transcriptional control, *Leishmania* uses several adaptive mechanisms to regulate gene expression when facing changing environmental conditions during its development. 3′ untranslated regions (3′ UTRs) were shown to be major players in monitoring mRNA stability and translation rates in this parasite[47–53]. To overcome stressful conditions like drug pressure, *Leishmania* also often relies on DNA copy number variations (CNVs) (aneuploidy, gene amplification, or gene deletion) for regulating the expression of drug targets, drug transporters, or other determinants of resistance. This is not restricted to *Leishmania*, however, and variations in gene dosage or chromosome copy numbers also influence drug susceptibility, adaptability, and proliferation in fungi and cancer cells[54–57]. In addition to CNVs, single-nucleotide polymorphisms (SNPs) in drug targets or in transporters can lead to drug resistance without the need for altering gene expression.

Copy number variations

During the last few decades, *Leishmania* parasites were considered to be essentially diploid but recent data have shown that aneuploidy seems to be the norm[58–64]. Within populations of *Leishmania* parasites, distinct aneuploidy patterns were shown to occur at the level of individual cells. This phenomenon was called mosaic aneuploidy and can translate into a seemingly average diploid population when the cumulative ploidy is derived from next-generation DNA sequencing data but in which few parasites actually share the same ploidy for individual chromosomes[62,63,65]. Interestingly, variations in the size and content of chromosomes have also been observed between different strains of the related trypanosomatid parasite *Trypanosoma cruzi*[66,67]. In the case of *Leishmania*, circumstantial links between the presence of supernumerary chromosomes or chromosomal losses and drug resistance have been observed[58–60,64,68–71], suggesting that a particular group of genes on the variant chromosomes may possibly act together in establishing resistance, but this has yet to be demonstrated.

Aneuploidy is generally linked to developmental abnormalities as best exemplified by the trisomy 21 syndrome in humans. However, *Leishmania* uses aneuploidy as a lifestyle. This is raising a number of questions about aneuploidy generation, stability, transmission, and biological significance (reviewed elsewhere[60,62,72]). In the absence of transcription initiation control, increases (or decreases) in chromosome copy number may serve as a strategy for regulating expression under environmental cues. This can happen at the level of whole chromosomes, and indeed there was a good correlation between chromosome ploidy and the level of DNA and RNA expression[59]. Increasing the copy numbers of entire chromosomes may lead to the overexpression of toxic genes, but the *Leishmania* genome (32 Mb) is spread in 34 to 36 chromosomes, thus reducing the co-expression of many genes. However, as explained in detail below, *Leishmania* also has the ability to amplify (or delete) specific smaller regions of DNA as part of extrachromosomal elements by recombination/rearrangements at the level of homologous repeated sequences (RSs). RNA levels derived from these amplifications are correlated to DNA copy number.

The genome of *Leishmania* is populated with repeated DNA sequences. A recent study highlighted the entire set of RSs in different *Leishmania* species, and it was found that the whole *Leishmania* genome has the potential to be rearranged at the level of those RSs for generating extrachromosomal elements[73]. Indeed, almost 2,000 RSs are distributed over the genome of *L. infantum* and these potentially support the formation of more than 3,000 extrachromosomal DNA elements[73]. Short interspersed degenerate retroposons (SIDERs) account for up to 65% of all RSs. SIDERs are truncated versions (~0.55 kb) of formerly active retroposons that are predominantly located in 3′ UTRs and have been associated with post-transcriptional regulation at the levels of both mRNA stability and translation[47,48,52,53,74,75]. Because SIDERs are degenerated, they were found in different RS groups. Remarkably, SIDER elements would have dual roles: one functional by regulating gene expression and a second one structural, providing the backbone to facilitate gene rearrangements for changing copy number of chromosomal DNA regions.

Extrachromosomal DNA amplifications are frequently detected in *Leishmania* parasites challenged with drugs or other stressful conditions[58,59,64,76–87]. The episomes are amplified as either circular or linear extrachromosomal DNA and formed through rearrangements at the level of direct or inverted homologous RSs, respectively (Figure 1)[73,80,88]. Interestingly, between 60% and 80% of the predicted amplicons appear to be already present in the population in the absence of selection and these pre-existing stochastic gene amplifications were shown to foster the selection of adaptive traits in response to drug pressure[73]. Beneficial amplicons were shown to increase in abundance upon higher drug pressure and to decrease when the drug is removed, allowing parasites to respond to a changing environment[73].

Since gene rearrangements through RSs are primary responses to drug pressure, a reasonable hypothesis was that identifying recombinase proteins involved in these rearrangements could lead to strategies to prevent the emergence of resistance. A first candidate was the RAD51 DNA repair protein, a key protein involved in homologous recombination (HR), a mechanism evolutionarily conserved in trypanosomatids[89]. Interestingly, the expression of RAD51

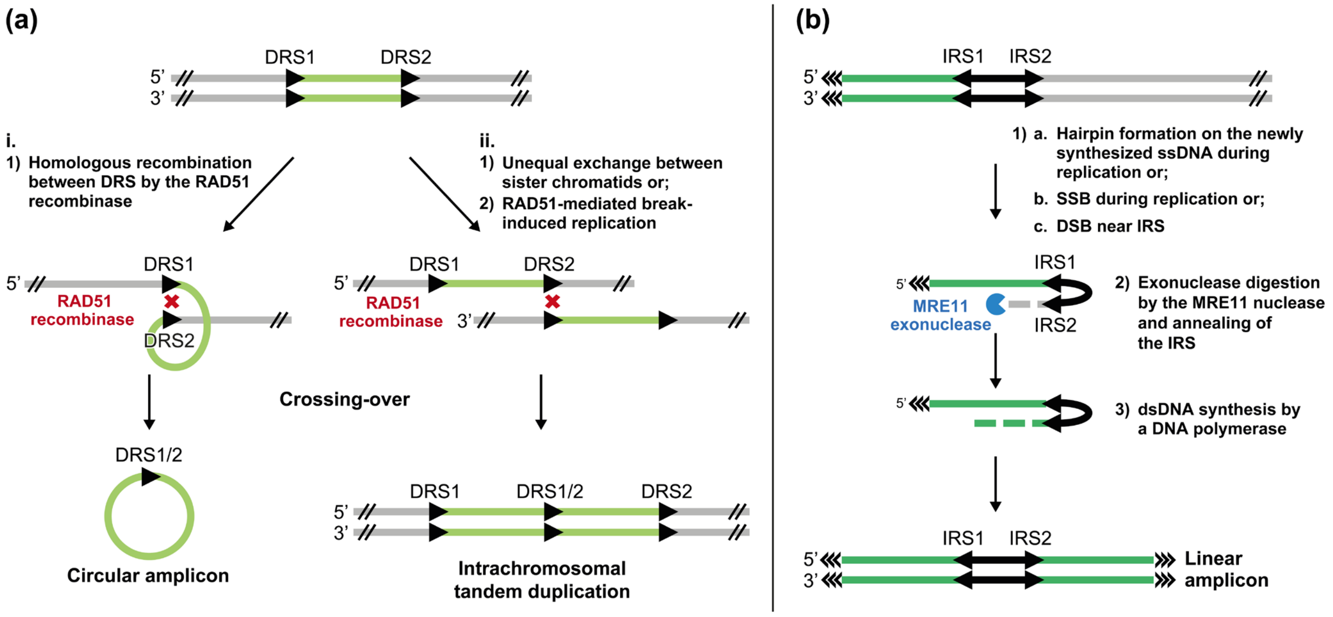

(a)
DRS1
DRS2
5′
3′
i.
1) Homologous recombination between DRS by the RAD51 recombinase
ii.
1) Unequal exchange between sister chromatids or;
2) RAD51-mediated break-induced replication
5′
DRS1
RAD51 recombinase
DRS2
5′
DRS1
DRS2
RAD51 recombinase
3′
Crossing-over
DRS1/2
Circular amplicon
DRS1
DRS1/2
DRS2
5′
3′
Intrachromosomal tandem duplication
(b)
IRS1
IRS2
5′
3′
1) a. Hairpin formation on the newly synthesized ssDNA during replication or;
b. SSB during replication or;
c. DSB near IRS
5′
IRS1
MRE11 exonuclease
IRS2
2) Exonuclease digestion by the MRE11 nuclease and annealing of the IRS
5′
3) dsDNA synthesis by a DNA polymerase
IRS1
IRS2
5′
3′
Linear amplicon

Figure 1. Potential mechanisms for gene amplification in *Leishmania*. (**a**) The RAD51 recombinase mediates homologous recombination between direct repeated sequences (DRS) and leads to (i) extrachromosomal circular amplicon or (ii) intrachromosomal tandem duplication by unequal sister chromatid exchange or RAD51-mediated break-induced replication. Black arrows represent DRS. (**b**) The MRE11 nuclease processes DNA ends after single-strand break (SSB), double-strand break (DSB), or hairpin formation during replication and leads to extrachromosomal linear amplification. Black arrows represent inverted repeated sequences (IRS). The green segments represent the amplified DNA regions. dsDNA, double-stranded DNA; ssDNA, single-stranded DNA.

was induced in *Leishmania* by DNA double-strand breaks (DSBs)[90,91]. Inactivating RAD51 led to viable parasites unable to generate circular extrachromosomal elements but still capable of producing linear amplicons upon drug pressure[73]. *Leishmania* has three RAD51 paralogs (RAD51-3, RAD51-4, and RAD51-6) that were shown to work as a complex in promoting HR through their capacity to stimulate RAD51 activity[92]. Inactivation of RAD51-4 was also shown to prevent the formation of circular amplicons in *L. infantum* exposed to drugs, but not of linear amplicons[92].

Linear amplicons are formed by the annealing of RSs found in an inverted orientation[64,73,93]. MRE11 is a DNA repair nuclease that interacts with RAD50 and NBS1 to form the MRN complex[94,95] and is important for DSB repair by HR[96,97] or for non-homologous end joining[89] (Figure 1b). Inactivation of MRE11 impaired the ability of *L. infantum* to form linear amplicons upon drug selection at the level of inverted RSs, although the capacity to generate circular amplicons was similar to that of wild-type parasites[93]. Moreover, a fully functional MRE11 is important for linear amplification, as parasites expressing DNA-binding-proficient but nuclease-deficient MRE11 exclusively generated circular amplicons during drug selection[93]. Interestingly, inactivation of MRE11 alone or along with its partner RAD50 led to extensive chromosomal translocation in *L. infantum*[98], showing that the MRE11/RAD50 complex is important for the maintenance of genome integrity in addition to its role in gene rearrangements. The number of enzymes involved in the formation of extrachromosomal elements suggests that targeting this pathway may not be a viable strategy for preventing the emergence of resistance, although this remains to be experimentally tested.

Single-nucleotide polymorphisms and small nucleotide insertions or deletions

Although CNVs are important contributors of drug resistance, SNPs and small nucleotide insertions or deletions (indels) can also contribute to resistance. This was proven with experimental drugs[99,100] and was highlighted with MTF where amino acid substitutions or non-sense mutations were observed in the MTF transporter (MT)[101] or in its Ros3 subunit[102]. This was further confirmed in additional mutants using whole genome sequencing[103,104] or by deep sequencing of MT[105]. Mutations detected in the MT gene of *L. infantum* isolates serially collected from an MTF-treated patient who had multiple relapses were shown to correlate with resistance[106], suggesting that MTF resistance could become a clinical reality in the near future.

Genome-wide surveys of genetic variations in *L. donovani* isolates from the Indian subcontinent supported the notion that resistance to antimonials emerged on several distinct occasions[58,107]. Isolates could also be clustered on the basis of their genetic structure and haplotypes, with some groups being enriched for non-responsive strains[58,107,108]. Interestingly, a particular group of highly resistant isolates that clustered together were found to share genomic features associated with resistance[107]. Among these were a higher copy number for the H-locus, coding for the well-characterized ABC transporter MRPA[109], and a homozygous two-base-pair insertion in the aquaglyceroporin 1 (AQP1) gene involved in Sb uptake and whose inactivation or downregulation is strongly correlated with resistance[71,107,110–117]. The potential for these genomic variants in predicting treatment outcome is exciting, given the lack of molecular markers for Sb resistance, but will need to be thoroughly evaluated by using larger and geographically diversified sets of well-defined isolates.

Exploiting copy number variations for understanding drug mode of action and resistance mechanisms

Target-based assays and phenotypic whole-cell-based assays are the cornerstones of drug discovery. The current trend for anti-parasitic agents is for whole cell assays. A drawback of phenotype-based assays is the lack of knowledge about the targets of hit compounds. Although the molecules could be brought to the clinic without further knowledge about their MOA, a clear understanding of the molecular targets will facilitate the improvement of a candidate drug through lead optimization. Characterization of drug-resistant mutants, which often revealed mutations or CNVs in drug targets or in proteins responsible for drug transport, is one strategy to pinpoint drug targets. However, it is salient to point out that this strategy has not yet led to targets against the current anti-leishmanials, although amplification of gene targets was observed in mutants made resistant to experimental drugs[64,99]. Since CNVs are often associated with resistance, forward genetic tools can experimentally mimic this. One such gain-of-function screen was based on functional cloning where *Leishmania* cosmid libraries were electroporated into *Leishmania* and these transfectants were selected for a specific phenotype[118]. Selection is possible because of the high copy number (and gene expression) of the cosmids. This screen was successfully applied while selecting for drug resistance or susceptibility[101,114,119–123]. This technique selects for cosmids conferring dominant phenotypes (leaving out less enriched cosmids) and is not easily amenable to high-throughput screening. The sensitivity of cosmid-based functional screening was enhanced by its recent coupling to next-generation sequencing in an approach termed Cos-Seq[124]. The proportion of parasites with cosmids providing a selective advantage is expected to rise with increasing drug pressure, and these can be tracked and quantified at each drug increment by Illumina sequencing[124,125]. Thus, the dynamics of cosmid enrichment can be followed over the entire course of selection instead of being monitored only at endpoint. Published or ongoing Cos-Seq screens using experimental drugs with known targets (for example, methotrexate, terbinafine, and 5-fluorouracil) confirmed the recovery of the relevant target genes by Cos-Seq[124]. Interestingly, Cos-Seq supported the hypothesis that the current anti-leishmanials (MF, AMB, Sb, PMM, and PTD) may not act via specific major protein targets[124]. Indeed, although an unprecedented number of resistance genes (known and novel) were isolated using Cos-Seq, none emerged as a clear target candidate and it is conceivable that these antiquated drugs are broadly cytotoxic by disrupting multiple minor targets. Whether some of the genes are genuine drug targets remains to be established, and non-protein targets represent another possibility as these would not be detected by Cos-Seq.

The advent of high-content screening for intracellular *L. donovani* amastigotes[126] is also key in the search for novel molecules having favourable anti-leishmanial properties directly on the intracellular stage of the parasite. This allowed the discovery of

192 new leads against visceral leishmaniasis from an initial set of 1.8 million compounds from GlaxoSmithKline[127]. An MOA could be hypothesized for 80 of the lead compounds using prior proprietary knowledge and bioinformatics analyses of TriTryp genomes, which revealed an over-representation of putative kinase inhibitors[127]. Cos-Seq was initially carried out with the insect form of the parasite but this could easily be adapted to intracellular parasites and this technique could be used to find the targets of these promising novel molecules or of other drugs repurposed against *Leishmania*[128].

The Cos-Seq technique does not allow the isolation of loss-of-function mutations such as those found in the aquaglyceroporin AQP1 or in the MT transporter genes (see above). These require high-throughput dominant negative screening approaches, like inducible RNA interference target sequencing (RIT-Seq), which proved instrumental in elucidating mechanisms of drug uptake in trypanosomes[129]. Although RNA interference is absent from the *L. Leishmania* subgenus, it is active in species of the *L. Viannia* subgenus[130]. The lack of inducible expression in *Leishmania* was also a limitation of this technique, but two recent reports have shown the feasibility of inducible expression in *Leishmania*[131,132]. Thus, it is theoretically possible to develop a technology similar to RIT-Seq in *L. Viannia* parasites. An alternative approach to RIT-Seq would be to rely on RNA-guided nuclease systems using clustered regularly interspaced short palindromic repeats (CRISPR) and CRISPR-associated (Cas) enzymes, as these have proven very efficient for achieving targeted genomic modifications in a wide range of genomes[133–135]. In trypanosomatid parasites, the CRISPR/Cas9 system derived from *Streptococcus pyogenes* has been used for disrupting genes in *L. major*[136], *L. donovani*[137], and *T. cruzi*[138] and in principle could be used for generating whole-genome Cas9-mediated gene deletion libraries.

Concluding remarks

The toolkit for drug target discovery and resistance mechanism elucidation for *Leishmania* is expanding. With new promising drug candidates in the pipeline and further technological developments, it should now be possible to find new targets which should further help in the control of this important neglected tropical disease.

Abbreviations

3′ UTR, 3′ untranslated region; AMB, amphotericin B; AQP1, aquaglyceroporin 1; Cas, clustered regularly interspaced short palindromic repeat-associated; CNV, copy number variation; CRISPR, clustered regularly interspaced short palindromic repeat; DSB, double-strand break; HR, homologous recombination; MOA, mode of action; mRNA, messenger RNA; MT, miltefosine transporter; MTF, miltefosine; PMM, paromomycin; PTD, pentamidine; RIT-Seq, RNA interference target sequencing; RS, repeated sequence; Sb, antimony; SIDER, short interspersed degenerate retroposon; SNP, single-nucleotide polymorphism.

Competing interests

The authors declare that they have no competing interests.

Grant information

BP is supported by grants from the Canadian Institutes of Health Research, the Natural Sciences and Engineering Research Council of Canada, and the Ministère du Developpement Économique de l'Innovation et de l'Exportation du Québec. MO is supported by grants from the Canadian Institutes of Health Research and holds a Canada Research Chair on Antimicrobial Resistance.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

F **F1000 recommended**

1. Alvar J, Vélez ID, Bern C, *et al.*: **Leishmaniasis worldwide and global estimates of its incidence.** *PLoS One.* 2012; **7**(5): e35671.
   PubMed Abstract | **Publisher Full Text** | **Free Full Text**
2. Bañuls AL, Hide M, Prugnolle F: ***Leishmania* and the leishmaniases: a parasite genetic update and advances in taxonomy, epidemiology and pathogenicity in humans.** *Adv Parasitol.* 2007; **64**: 1–109.
   PubMed Abstract | **Publisher Full Text**
3. Guerin PJ, Olliaro P, Sundar S, *et al.*: **Visceral leishmaniasis: current status of control, diagnosis, and treatment, and a proposed research and development agenda.** *Lancet Infect Dis.* 2002; **2**(8): 494–501.
   PubMed Abstract | **Publisher Full Text**
4. David CV, Craft N: **Cutaneous and mucocutaneous leishmaniasis.** *Dermatol Ther.* 2009; **22**(6): 491–502.
   PubMed Abstract | **Publisher Full Text**
5. Herwaldt BL: **Leishmaniasis.** *Lancet.* 1999; **354**(9185): 1191–9.
   PubMed Abstract | **Publisher Full Text**
6. Gradoni L: **Canine *Leishmania* vaccines: still a long way to go.** *Vet Parasitol.* 2015; **208**(1–2): 94–100.
   PubMed Abstract | **Publisher Full Text**
7. Handman E: **Leishmaniasis: current status of vaccine development.** *Clin Microbiol Rev.* 2001; **14**(2): 229–43.
   PubMed Abstract | **Publisher Full Text** | **Free Full Text**
8. Murray HW, Berman JD, Davies CR, *et al.*: **Advances in leishmaniasis.** *Lancet.* 2005; **366**(9496): 1561–77.
   PubMed Abstract | **Publisher Full Text**
9. Palatnik-de-Sousa CB: **Vaccines for leishmaniasis in the fore coming 25 years.** *Vaccine.* 2008; **26**(14): 1709–24.
   PubMed Abstract | **Publisher Full Text**
10. F Sundar S, Chakravarty J, Agarwal D, *et al.*: **Single-dose liposomal amphotericin B for visceral leishmaniasis in India.** *N Engl J Med.* 2010; **362**(6): 504–12.
    PubMed Abstract | **Publisher Full Text** | **F1000 Recommendation**
11. Bern C, Adler-Moore J, Berenguer J, *et al.*: **Liposomal amphotericin B for the treatment of visceral leishmaniasis.** *Clin Infect Dis.* 2006; **43**(7): 917–24.
    PubMed Abstract | **Publisher Full Text**
12. Berman JD, Badaro R, Thakur CP, *et al.*: **Efficacy and safety of liposomal amphotericin B (AmBisome) for visceral leishmaniasis in endemic developing countries.** *Bull World Health Organ.* 1998; **76**(1): 25–32.
    PubMed Abstract | **Free Full Text**
13. Lachaud L, Bourgeois N, Plourde M, *et al.*: **Parasite susceptibility to amphotericin B in failures of treatment for visceral leishmaniasis in patients coinfected with HIV type 1 and *Leishmania infantum*.** *Clin Infect Dis.* 2009; **48**(2): e16–22.
    PubMed Abstract | **Publisher Full Text**

14. Croft SL, Neal RA, Pendergast W, *et al.*: **The activity of alkyl phosphorylcholines and related derivatives against *Leishmania donovani*.** *Biochem Pharmacol.* 1987; **36**(16): 2633–6.
PubMed Abstract | Publisher Full Text

15. Jha TK, Sundar S, Thakur CP, *et al.*: **Miltefosine, an oral agent, for the treatment of Indian visceral leishmaniasis.** *N Engl J Med.* 1999; **341**(24): 1795–800.
PubMed Abstract | Publisher Full Text

16. Sundar S, Jha TK, Thakur CP, *et al.*: **Oral miltefosine for Indian visceral leishmaniasis.** *N Engl J Med.* 2002; **347**(22): 1739–46.
PubMed Abstract | Publisher Full Text

17. F Burza S, Nabi E, Mahajan R, *et al.*: **One-year follow-up of immunocompetent male patients treated with miltefosine for primary visceral leishmaniasis in Bihar, India.** *Clin Infect Dis.* 2013; **57**(9): 1363–4.
PubMed Abstract | Publisher Full Text | F1000 Recommendation

18. F Sundar S, Singh A, Rai M, *et al.*: **Efficacy of miltefosine in the treatment of visceral leishmaniasis in India after a decade of use.** *Clin Infect Dis.* 2012; **55**(4): 543–50.
PubMed Abstract | Publisher Full Text | F1000 Recommendation

19. Pandey BD, Pandey K, Kaneko O, *et al.*: **Relapse of visceral leishmaniasis after miltefosine treatment in a Nepalese patient.** *Am J Trop Med Hyg.* 2009; **80**(4): 580–2.
PubMed Abstract

20. F Rijal S, Ostyn B, Uranw S, *et al.*: **Increasing failure of miltefosine in the treatment of Kala-azar in Nepal and the potential role of parasite drug resistance, reinfection, or noncompliance.** *Clin Infect Dis.* 2013; **56**(11): 1530–8.
PubMed Abstract | Publisher Full Text | F1000 Recommendation

21. Jha TK, Olliaro P, Thakur CP, *et al.*: **Randomised controlled trial of aminosidine (paromomycin) v sodium stibogluconate for treating visceral leishmaniasis in North Bihar, India.** *BMJ.* 1998; **316**(7139): 1200–5.
PubMed Abstract | Publisher Full Text | Free Full Text

22. Sundar S, Jha TK, Thakur CP, *et al.*: **Injectable paromomycin for Visceral leishmaniasis in India.** *N Engl J Med.* 2007; **356**(25): 2571–81.
PubMed Abstract | Publisher Full Text

23. Hailu A, Musa A, Wasunna M, *et al.*: **Geographical variation in the response of visceral leishmaniasis to paromomycin in East Africa: a multicentre, open-label, randomized trial.** *PLoS Negl Trop Dis.* 2010; **4**(10): e709.
PubMed Abstract | Publisher Full Text | Free Full Text

24. Musa AM, Younis B, Fadlalla A, *et al.*: **Paromomycin for the treatment of visceral leishmaniasis in Sudan: a randomized, open-label, dose-finding study.** *PLoS Negl Trop Dis.* 2010; **4**(10): e855.
PubMed Abstract | Publisher Full Text | Free Full Text

25. Lai A Fat EJ, Vrede MA, Soetosenojo RM, *et al.*: **Pentamidine, the drug of choice for the treatment of cutaneous leishmaniasis in Surinam.** *Int J Dermatol.* 2002; **41**(11): 796–800.
PubMed Abstract | Publisher Full Text

26. Roussel M, Nacher M, Frémont G, *et al.*: **Comparison between one and two injections of pentamidine isethionate, at 7 mg/kg in each injection, in the treatment of cutaneous leishmaniasis in French Guiana.** *Ann Trop Med Parasitol.* 2006; **100**(4): 307–14.
PubMed Abstract | Publisher Full Text

27. Soto J, Buffet P, Grogl M, *et al.*: **Successful treatment of Colombian cutaneous leishmaniasis with four injections of pentamidine.** *Am J Trop Med Hyg.* 1994; **50**(1): 107–11.
PubMed Abstract

28. Mandal G, Wyllie S, Singh N, *et al.*: **Increased levels of thiols protect antimony unresponsive *Leishmania donovani* field isolates against reactive oxygen species generated by trivalent antimony.** *Parasitology.* 2007; **134**(Pt 12): 1679–87.
PubMed Abstract | Publisher Full Text | Free Full Text

29. Mehta A, Shaha C: **Mechanism of metalloid-induced death in *Leishmania* spp.: role of iron, reactive oxygen species, Ca^{2+}, and glutathione.** *Free Radic Biol Med.* 2006; **40**(10): 1857–68.
PubMed Abstract | Publisher Full Text

30. Mookerjee Basu J, Mookerjee A, Sen P, *et al.*: **Sodium antimony gluconate induces generation of reactive oxygen species and nitric oxide via phosphoinositide 3-kinase and mitogen-activated protein kinase activation in *Leishmania donovani*-infected macrophages.** *Antimicrob Agents Chemother.* 2006; **50**(5): 1788–97.
PubMed Abstract | Publisher Full Text | Free Full Text

31. Moreira W, Leprohon P, Ouellette M: **Tolerance to drug-induced cell death favours the acquisition of multidrug resistance in *Leishmania*.** *Cell Death Dis.* 2011; **2**: e201.
PubMed Abstract | Publisher Full Text | Free Full Text

32. Wyllie S, Cunningham ML, Fairlamb AH: **Dual action of antimonial drugs on thiol redox metabolism in the human pathogen *Leishmania donovani*.** *J Biol Chem.* 2004; **279**(38): 39925–32.
PubMed Abstract | Publisher Full Text

33. Lee N, Bertholet S, Debrabant A, *et al.*: **Programmed cell death in the unicellular protozoan parasite *Leishmania*.** *Cell Death Differ.* 2002; **9**(1): 53–64.
PubMed Abstract | Publisher Full Text

34. Sereno D, Holzmuller P, Mangot I, *et al.*: **Antimonial-mediated DNA fragmentation in *Leishmania infantum* amastigotes.** *Antimicrob Agents Chemother.* 2001; **45**(7): 2064–9.
PubMed Abstract | Publisher Full Text | Free Full Text

35. Sudhandiran G, Shaha C: **Antimonial-induced increase in intracellular Ca^{2+} through non-selective cation channels in the host and the parasite is responsible for apoptosis of intracellular *Leishmania donovani* amastigotes.** *J Biol Chem.* 2003; **278**(27): 25120–32.
PubMed Abstract | Publisher Full Text

36. Vergnes B, Gourbal B, Girard I, *et al.*: **A proteomics screen implicates HSP83 and a small kinetoplastid calpain-related protein in drug resistance in *Leishmania donovani* clinical field isolates by modulating drug-induced programmed cell death.** *Mol Cell Proteomics.* 2007; **6**(1): 88–101.
PubMed Abstract | Publisher Full Text

37. F Ivens AC, Peacock CS, Worthey EA, *et al.*: **The genome of the kinetoplastid parasite, *Leishmania major*.** *Science.* 2005; **309**(5733): 436–42.
PubMed Abstract | Publisher Full Text | Free Full Text | F1000 Recommendation

38. F Peacock CS, Seeger K, Harris D, *et al.*: **Comparative genomic analysis of three *Leishmania* species that cause diverse human disease.** *Nat Genet.* 2007; **39**(7): 839–47.
PubMed Abstract | Publisher Full Text | Free Full Text | F1000 Recommendation

39. Britto C, Ravel C, Bastien P, *et al.*: **Conserved linkage groups associated with large-scale chromosomal rearrangements between Old World and New World *Leishmania* genomes.** *Gene.* 1998; **222**(1): 107–17.
PubMed Abstract | Publisher Full Text

40. Raymond F, Boisvert S, Roy G, *et al.*: **Genome sequencing of the lizard parasite *Leishmania tarentolae* reveals loss of genes associated to the intracellular stage of human pathogenic species.** *Nucleic Acids Res.* 2012; **40**(3): 1131–47.
PubMed Abstract | Publisher Full Text | Free Full Text

41. Wincker P, Ravel C, Blaineau C, *et al.*: **The *Leishmania* genome comprises 36 chromosomes conserved across widely divergent human pathogenic species.** *Nucleic Acids Res.* 1996; **24**(9): 1688–94.
PubMed Abstract | Publisher Full Text | Free Full Text

42. El-Sayed NM, Myler PJ, Blandin G, *et al.*: **Comparative genomics of trypanosomatid parasitic protozoa.** *Science.* 2005; **309**(5733): 404–9.
PubMed Abstract | Publisher Full Text

43. Martinez-Calvillo S, Yan S, Nguyen D, *et al.*: **Transcription of *Leishmania major* Friedlin chromosome 1 initiates in both directions within a single region.** *Mol Cell.* 2003; **11**(5): 1291–9.
PubMed Abstract | Publisher Full Text

44. Martinez-Calvillo S, Nguyen D, Stuart K, *et al.*: **Transcription initiation and termination on *Leishmania major* chromosome 3.** *Eukaryotic Cell.* 2004; **3**(2): 506–17.
PubMed Abstract | Publisher Full Text | Free Full Text

45. Haile S, Papadopoulou B: **Developmental regulation of gene expression in trypanosomatid parasitic protozoa.** *Curr Opin Microbiol.* 2007; **10**(6): 569–77.
PubMed Abstract | Publisher Full Text

46. Matthews KR, Tschudi C, Ullu E: **A common pyrimidine-rich motif governs trans-splicing and polyadenylation of tubulin polycistronic pre-mRNA in trypanosomes.** *Genes Dev.* 1994; **8**(4): 491–501.
PubMed Abstract | Publisher Full Text

47. Boucher N, Wu Y, Dumas C, *et al.*: **A common mechanism of stage-regulated gene expression in *Leishmania* mediated by a conserved 3'-untranslated region element.** *J Biol Chem.* 2002; **277**(22): 19511–20.
PubMed Abstract | Publisher Full Text

48. F Bringaud F, Muller M, Cerqueira GC, *et al.*: **Members of a large retroposon family are determinants of post-transcriptional gene expression in *Leishmania*.** *PLoS Pathog.* 2007; **3**(9): 1291–307, e136.
PubMed Abstract | Publisher Full Text | Free Full Text | F1000 Recommendation

49. Dupe A, Dumas C, Papadopoulou B: **An Alba-domain protein contributes to the stage-regulated stability of amastin transcripts in *Leishmania*.** *Mol Microbiol.* 2014; **91**(3): 548–61.
PubMed Abstract | Publisher Full Text

50. Haile S, Dupe A, Papadopoulou B: **Deadenylation-independent stage-specific mRNA degradation in *Leishmania*.** *Nucleic Acids Res.* 2008; **36**(5): 1634–44.
PubMed Abstract | Publisher Full Text | Free Full Text

51. McNicoll F, Muller M, Cloutier S, *et al.*: **Distinct 3'-untranslated region elements regulate stage-specific mRNA accumulation and translation in *Leishmania*.** *J Biol Chem.* 2005; **280**(42): 35238–46.
PubMed Abstract | Publisher Full Text

52. Muller M, Padmanabhan PK, Papadopoulou B: **Selective inactivation of SIDER2 retroposon-mediated mRNA decay contributes to stage- and species-specific gene expression in *Leishmania*.** *Mol Microbiol.* 2010; **77**(2): 471–91.
PubMed Abstract | Publisher Full Text

53. Muller M, Padmanabhan PK, Rochette A, *et al.*: **Rapid decay of unstable *Leishmania* mRNAs bearing a conserved retroposon signature 3'-UTR motif is initiated by a site-specific endonucleolytic cleavage without prior deadenylation.** *Nucleic Acids Res.* 2010; **38**(17): 5867–83.
PubMed Abstract | Publisher Full Text | Free Full Text

54. Gordon DJ, Resio B, Pellman D: **Causes and consequences of aneuploidy in cancer.** *Nat Rev Genet.* 2012; **13**(3): 189–203.
PubMed Abstract | Publisher Full Text

55. Pfau SJ, Amon A: **Chromosomal instability and aneuploidy in cancer: from yeast to man.** *EMBO Rep.* 2012; **13**(6): 515–27.
PubMed Abstract | Publisher Full Text | Free Full Text

56. Rutledge SD, Cimini D: **Consequences of aneuploidy in sickness and in health.** *Curr Opin Cell Biol.* 2016; **40**: 41–6.
PubMed Abstract | Publisher Full Text

57. Shor E, Perlin DS: **Coping with stress and the emergence of multidrug resistance in fungi.** *PLoS Pathog.* 2015; **11**(3): e1004668.
PubMed Abstract | Publisher Full Text | Free Full Text

58. F Downing T, Imamura H, Decuypere S, *et al.*: **Whole genome sequencing of multiple *Leishmania donovani* clinical isolates provides insights into population structure and mechanisms of drug resistance.** *Genome Res.* 2011; **21**(12): 2143–56.
PubMed Abstract | Publisher Full Text | Free Full Text | F1000 Recommendation

59. Leprohon P, Legare D, Raymond F, *et al.*: **Gene expression modulation is associated with gene amplification, supernumerary chromosomes and chromosome loss in antimony-resistant *Leishmania infantum*.** *Nucleic Acids Res.* 2009; **37**(5): 1387–99.
PubMed Abstract | Publisher Full Text | Free Full Text

60. Mannaert A, Downing T, Imamura H, *et al.*: **Adaptive mechanisms in pathogens: universal aneuploidy in *Leishmania*.** *Trends Parasitol.* 2012; **28**(9): 370–6.
PubMed Abstract | Publisher Full Text

61. F Rogers MB, Hilley JD, Dickens NJ, *et al.*: **Chromosome and gene copy number variation allow major structural change between species and strains of *Leishmania*.** *Genome Res.* 2011; **21**(12): 2129–42.
PubMed Abstract | Publisher Full Text | Free Full Text | F1000 Recommendation

62. Sterkers Y, Lachaud L, Bourgeois N, *et al.*: **Novel insights into genome plasticity in Eukaryotes: mosaic aneuploidy in *Leishmania*.** *Mol Microbiol.* 2012; **86**(1): 15–23.
PubMed Abstract | Publisher Full Text

63. Sterkers Y, Lachaud L, Crobu L, *et al.*: **FISH analysis reveals aneuploidy and continual generation of chromosomal mosaicism in *Leishmania major*.** *Cell Microbiol.* 2011; **13**(2): 274–83.
PubMed Abstract | Publisher Full Text

64. Ubeda JM, Légaré D, Raymond F, *et al.*: **Modulation of gene expression in drug resistant *Leishmania* is associated with gene amplification, gene deletion and chromosome aneuploidy.** *Genome Biol.* 2008; **9**(7): R115.
PubMed Abstract | Publisher Full Text | Free Full Text

65. F Lachaud L, Bourgeois N, Kuk N, *et al.*: **Constitutive mosaic aneuploidy is a unique genetic feature widespread in the *Leishmania* genus.** *Microbes Infect.* 2014; **16**(1): 61–6.
PubMed Abstract | Publisher Full Text | F1000 Recommendation

66. Minning TA, Weatherly DB, Flibotte S, *et al.*: **Widespread, focal copy number variations (CNV) and whole chromosome aneuploidies in *Trypanosoma cruzi* strains revealed by array comparative genomic hybridization.** *BMC Genomics.* 2011; **12**: 139.
PubMed Abstract | Publisher Full Text | Free Full Text

67. F Reis-Cunha JL, Rodrigues-Luiz GF, Valdivia HO, *et al.*: **Chromosomal copy number variation reveals differential levels of genomic plasticity in distinct *Trypanosoma cruzi* strains.** *BMC Genomics.* 2015; **16**: 499.
PubMed Abstract | Publisher Full Text | Free Full Text | F1000 Recommendation

68. Brotherton MC, Bourassa S, Leprohon P, *et al.*: **Proteomic and genomic analyses of antimony resistant *Leishmania infantum* mutant.** *PLoS One.* 2013; **8**(11): e81899.
PubMed Abstract | Publisher Full Text | Free Full Text

69. do Monte-Neto RL, Coelho AC, Raymond F, *et al.*: **Gene expression profiling and molecular characterization of antimony resistance in *Leishmania amazonensis*.** *PLoS Negl Trop Dis.* 2011; **5**(5): e1167.
PubMed Abstract | Publisher Full Text | Free Full Text

70. Kumar P, Lodge R, Raymond F, *et al.*: **Gene expression modulation and the molecular mechanisms involved in Nelfinavir resistance in *Leishmania donovani* axenic amastigotes.** *Mol Microbiol.* 2013; **89**(3): 565–82.
PubMed Abstract | Publisher Full Text

71. Mukherjee A, Boisvert S, Monte-Neto RL, *et al.*: **Telomeric gene deletion and intrachromosomal amplification in antimony-resistant *Leishmania*.** *Mol Microbiol.* 2013; **88**(1): 189–202.
PubMed Abstract | Publisher Full Text

72. F Sterkers Y, Crobu L, Lachaud L, *et al.*: **Parasexuality and mosaic aneuploidy in *Leishmania*: alternative genetics.** *Trends Parasitol.* 2014; **30**(9): 429–35.
PubMed Abstract | Publisher Full Text | F1000 Recommendation

73. Ubeda JM, Raymond F, Mukherjee A, *et al.*: **Genome-wide stochastic adaptive DNA amplification at direct and inverted DNA repeats in the parasite *Leishmania*.** *PLoS Biol.* 2014; **12**(5): e1001868.
PubMed Abstract | Publisher Full Text | Free Full Text

74. Ouellette M, Papadopoulou B: **Coordinated gene expression by post-transcriptional regulons in African trypanosomes.** *J Biol.* 2009; **8**(11): 100.
PubMed Abstract | Publisher Full Text | Free Full Text

75. Smith M, Bringaud F, Papadopoulou B: **Organization and evolution of two SIDER retroposon subfamilies and their impact on the *Leishmania* genome.** *BMC Genomics.* 2009; **10**: 240.
PubMed Abstract | Publisher Full Text | Free Full Text

76. Beverley SM: **Gene amplification in *Leishmania*.** *Annu Rev Microbiol.* 1991; **45**: 417–44.
PubMed Abstract | Publisher Full Text

77. Beverley SM, Coderre JA, Santi DV, *et al.*: **Unstable DNA amplifications in methotrexate-resistant Leishmania consist of extrachromosomal circles which relocalize during stabilization.** *Cell.* 1984; **38**(2): 431–9.
PubMed Abstract | Publisher Full Text

78. Garvey EP, Santi DV: **Stable amplified DNA in drug-resistant Leishmania exists as extrachromosomal circles.** *Science.* 1986; **233**(4763): 535–40.
PubMed Abstract | Publisher Full Text

79. Grondin K, Haimeur A, Mukhopadhyay R, *et al.*: **Co-amplification of the gamma-glutamylcysteine synthetase gene *gsh1* and of the ABC transporter gene *pgpA* in arsenite-*resistant Leishmania tarentolae*.** *EMBO J.* 1997; **16**(11): 3057–65.
PubMed Abstract | Publisher Full Text | Free Full Text

80. Grondin K, Papadopoulou B, Ouellette M: **Homologous recombination between direct repeat sequences yields P-glycoprotein containing amplicons in arsenite resistant *Leishmania*.** *Nucleic Acids Res.* 1993; **21**(8): 1895–901.
PubMed Abstract | Publisher Full Text | Free Full Text

81. Haimeur A, Brochu C, Genest P, *et al.*: **Amplification of the ABC transporter gene *PGPA* and increased trypanothione levels in potassium antimonyl tartrate (SbIII) resistant *Leishmania tarentolae*.** *Mol Biochem Parasitol.* 2000; **108**(1): 131–5.
PubMed Abstract | Publisher Full Text

82. Kundig C, Leblanc E, Papadopoulou B, *et al.*: **Role of the locus and of the resistance gene on gene amplification frequency in methotrexate resistant *Leishmania tarentolae*.** *Nucleic Acids Res.* 1999; **27**(18): 3653–9.
PubMed Abstract | Publisher Full Text | Free Full Text

83. Mittal MK, Rai S, Ashutosh, *et al.*: **Characterization of natural antimony resistance in *Leishmania donovani isolates*.** *Am J Trop Med Hyg.* 2007; **76**(4): 681–8.
PubMed Abstract

84. Mukherjee A, Padmanabhan PK, Singh S, *et al.*: **Role of ABC transporter MRPA, gamma-glutamylcysteine synthetase and ornithine decarboxylase in natural antimony-resistant isolates of *Leishmania donovani*.** *J Antimicrob Chemother.* 2007; **59**(2): 204–11.
PubMed Abstract | Publisher Full Text

85. Ouellette M, Hettema E, Wüst D, *et al.*: **Direct and inverted DNA repeats associated with P-glycoprotein gene amplification in drug resistant Leishmania.** *EMBO J.* 1991; **10**(4): 1009–16.
PubMed Abstract | Free Full Text

86. Papadopoulou B, Roy G, Ouellette M: **Frequent amplification of a short chain dehydrogenase gene as part of circular and linear amplicons in methotrexate resistant *Leishmania*.** *Nucleic Acids Res.* 1993; **21**(18): 4305–12.
PubMed Abstract | Free Full Text

87. White TC, Fase-Fowler F, van Luenen H, *et al.*: **The H circles of *Leishmania tarentolae* are a unique amplifiable system of oligomeric DNAs associated with drug resistance.** *J Biol Chem.* 1988; **263**(32): 16977–83.
PubMed Abstract

88. Navarro M, Liu J, Muthui D, *et al.*: **Inverted repeat structure and homologous sequences in the LD_1 amplicons of Leishmania spp.** *Mol Biochem Parasitol.* 1994; **68**(1): 69–80.
PubMed Abstract | Publisher Full Text

89. Genois MM, Paquet ER, Laffitte MC, *et al.*: **DNA repair pathways in trypanosomatids: from DNA repair to drug resistance.** *Microbiol Mol Biol Rev.* 2014; **78**(1): 40–73.
PubMed Abstract | Publisher Full Text | Free Full Text

90. Genois MM, Mukherjee A, Ubeda JM, *et al.*: **Interactions between BRCA2 and RAD51 for promoting homologous recombination in *Leishmania infantum*.** *Nucleic Acids Res.* 2012; **40**(14): 6570–84.
PubMed Abstract | Publisher Full Text | Free Full Text

91. McKean PG, Keen JK, Smith DF, *et al.*: **Identification and characterisation of a *RAD51* gene from *Leishmania major*.** *Mol Biochem Parasitol.* 2001; **115**(2): 209–16.
PubMed Abstract | Publisher Full Text

92. Genois MM, Plourde M, Éthier C, *et al.*: **Roles of Rad51 paralogs for promoting homologous recombination in *Leishmania infantum*.** *Nucleic Acids Res.* 2015; **43**(5): 2701–15.
PubMed Abstract | Publisher Full Text | Free Full Text

93. Laffitte MC, Genois MM, Mukherjee A, *et al.*: **Formation of linear amplicons with inverted duplications in *Leishmania* requires the MRE11 nuclease.** *PLoS Genet.* 2014; **10**(12): e1004805.
PubMed Abstract | Publisher Full Text | Free Full Text

94. Assenmacher N, Hopfner KP: **MRE11/RAD50/NBS1: complex activities.** *Chromosoma.* 2004; **113**(4): 157–66.
PubMed Abstract | Publisher Full Text

95. Stracker TH, Petrini JH: **The MRE11 complex: starting from the ends.** *Nat Rev Mol Cell Biol.* 2011; **12**(2): 90–103.
PubMed Abstract | Publisher Full Text | Free Full Text

96. F Shibata A, Moiani D, Arvai AS, *et al.*: **DNA double-strand break repair pathway choice is directed by distinct MRE11 nuclease activities.** *Mol Cell.* 2014; **53**(1): 7–18.
PubMed Abstract | Publisher Full Text | Free Full Text | F1000 Recommendation

97. Mimitou EP, Symington LS: **DNA end resection--unraveling the tail.** *DNA Repair (Amst).* 2011; **10**(3): 344–8.
PubMed Abstract | Publisher Full Text | Free Full Text

98. Laffitte MC, Leprohon P, Hainse M, *et al.*: **Chromosomal Translocations in the Parasite *Leishmania* by a MRE11/RAD50-Independent Microhomology-Mediated End Joining Mechanism.** *PLoS Genet.* 2016; **12**(6): e1006117.
PubMed Abstract | Publisher Full Text | Free Full Text

99. Ritt JF, Raymond F, Leprohon P, *et al.*: **Gene amplification and point mutations**

in pyrimidine metabolic genes in 5-fluorouracil resistant *Leishmania infantum*. *PLoS Negl Trop Dis.* 2013; **7**(11): e2564.
PubMed Abstract | Publisher Full Text | Free Full Text

100. Vasudevan G, Ullman B, Landfear SM: **Point mutations in a nucleoside transporter gene from *Leishmania donovani* confer drug resistance and alter substrate selectivity.** *Proc Natl Acad Sci USA.* 2001; **98**(11): 6092–7.
PubMed Abstract | Publisher Full Text | Free Full Text

101. Pérez-Victoria FJ, Gamarro F, Ouellette M, *et al.*: **Functional cloning of the miltefosine transporter. A novel P-type phospholipid translocase from *Leishmania* involved in drug resistance.** *J Biol Chem.* 2003; **278**(50): 49965–71.
PubMed Abstract | Publisher Full Text

102. Pérez-Victoria FJ, Sánchez-Cañete MP, Castanys S, *et al.*: **Phospholipid translocation and miltefosine potency require both *L. donovani* miltefosine transporter and the new protein LdRos3 in *Leishmania* parasites.** *J Biol Chem.* 2006; **281**(33): 23766–75.
PubMed Abstract | Publisher Full Text

103. Coelho AC, Boisvert S, Mukherjee A, *et al.*: **Multiple mutations in heterogeneous miltefosine-resistant *Leishmania major* population as determined by whole genome sequencing.** *PLoS Negl Trop Dis.* 2012; **6**(2): e1512.
PubMed Abstract | Publisher Full Text | Free Full Text

104. F Mondelaers A, Sanchez-Cañete MP, Hendrickx S, *et al.*: **Genomic and Molecular Characterization of Miltefosine Resistance in *Leishmania infantum* Strains with Either Natural or Acquired Resistance through Experimental Selection of Intracellular Amastigotes.** *PLoS One.* 2016; **11**(4): e0154101.
PubMed Abstract | Publisher Full Text | Free Full Text | F1000 Recommendation

105. Laffitte MN, Leprohon P, Légaré D, *et al.*: **Deep-sequencing revealing mutation dynamics in the miltefosine transporter gene in *Leishmania infantum* selected for miltefosine resistance.** *Parasitol Res.* 2016; 1–5.
PubMed Abstract | Publisher Full Text

106. Cojean S, Houzé S, Haouchine D, *et al.*: ***Leishmania* resistance to miltefosine associated with genetic marker.** *Emerging Infect Dis.* 2012; **18**(4): 704–6.
PubMed Abstract | Publisher Full Text | Free Full Text

107. F Imamura H, Downing T, Van den Broeck F, *et al.*: **Evolutionary genomics of epidemic visceral leishmaniasis in the Indian subcontinent.** *eLife.* 2016; **5**: pii: e12613.
PubMed Abstract | Publisher Full Text | Free Full Text | F1000 Recommendation

108. Vanaerschot M, Decuypere S, Downing T, *et al.*: **Genetic markers for SSG resistance in *Leishmania donovani* and SSG treatment failure in visceral leishmaniasis patients of the Indian subcontinent.** *J Infect Dis.* 2012; **206**(5): 752–5.
PubMed Abstract | Publisher Full Text | Free Full Text

109. Légaré D, Richard D, Mukhopadhyay R, *et al.*: **The *Leishmania* ATP-binding cassette protein PGPA is an intracellular metal-thiol transporter ATPase.** *J Biol Chem.* 2001; **276**(28): 26301–7.
PubMed Abstract | Publisher Full Text

110. Decuypere S, Rijal S, Yardley V, *et al.*: **Gene expression analysis of the mechanism of natural Sb(V) resistance in *Leishmania donovani* isolates from Nepal.** *Antimicrob Agents Chemother.* 2005; **49**(11): 4616–21.
PubMed Abstract | Publisher Full Text | Free Full Text

111. Gourbal B, Sonuc N, Bhattacharjee H, *et al.*: **Drug uptake and modulation of drug resistance in *Leishmania* by an aquaglyceroporin.** *J Biol Chem.* 2004; **279**(30): 31010–7.
PubMed Abstract | Publisher Full Text

112. Mandal G, Mandal S, Sharma M, *et al.*: **Species-specific antimonial sensitivity in *Leishmania* is driven by post-transcriptional regulation of AQP1.** *PLoS Negl Trop Dis.* 2015; **9**(2): e0003500.
PubMed Abstract | Publisher Full Text | Free Full Text

113. Mandal S, Maharjan M, Singh S, *et al.*: **Assessing aquaglyceroporin gene status and expression profile in antimony-susceptible and -resistant clinical isolates of *Leishmania donovani* from India.** *J Antimicrob Chemother.* 2010; **65**(3): 496–507.
PubMed Abstract | Publisher Full Text

114. Marquis N, Gourbal B, Rosen BP, *et al.*: **Modulation in aquaglyceroporin *AQP1* gene transcript levels in drug-resistant *Leishmania*.** *Mol Microbiol.* 2005; **57**(6): 1690–9.
PubMed Abstract | Publisher Full Text

115. Monte-Neto R, Laffitte MC, Leprohon P, *et al.*: **Intrachromosomal amplification, locus deletion and point mutation in the aquaglyceroporin AQP1 gene in antimony resistant *Leishmania (Viannia)* guyanensis.** *PLoS Negl Trop Dis.* 2015; **9**: e0003476.
PubMed Abstract | Publisher Full Text | Free Full Text

116. Plourde M, Ubeda JM, Mandal G, *et al.*: **Generation of an aquaglyceroporin *AQP1* null mutant in *Leishmania major*.** *Mol Biochem Parasitol.* 2015; **201**(2): 108–11.
PubMed Abstract | Publisher Full Text

117. Uzcategui NL, Zhou Y, Figarella K, *et al.*: **Alteration in glycerol and metalloid permeability by a single mutation in the extracellular C-loop of *Leishmania major* aquaglyceroporin LmAQP1.** *Mol Microbiol.* 2008; **70**(6): 1477–86.
PubMed Abstract | Publisher Full Text | Free Full Text

118. Ryan KA, Garraway LA, Descoteaux A, *et al.*: **Isolation of virulence genes directing surface glycosyl-phosphatidylinositol synthesis by functional complementation of *Leishmania*.** *Proc Natl Acad Sci USA.* 1993; **90**(18): 8609–13.
PubMed Abstract | Publisher Full Text | Free Full Text

119. Carter NS, Drew ME, Sanchez M, *et al.*: **Cloning of a novel inosine-guanosine transporter gene from *Leishmania donovani* by functional rescue of a transport-deficient mutant.** *J Biol Chem.* 2000; **275**(27): 20935–41.
PubMed Abstract | Publisher Full Text

120. Coelho AC, Beverley SM, Cotrim PC: **Functional genetic identification of PRP1, an ABC transporter superfamily member conferring pentamidine resistance in *Leishmania major*.** *Mol Biochem Parasitol.* 2003; **130**(2): 83–90.
PubMed Abstract | Publisher Full Text

121. Cotrim PC, Garrity LK, Beverley SM: **Isolation of genes mediating resistance to inhibitors of nucleoside and ergosterol metabolism in *Leishmania* by overexpression/selection.** *J Biol Chem.* 1999; **274**(53): 37723–30.
PubMed Abstract | Publisher Full Text

122. Kundig C, Haimeur A, Legare D, *et al.*: **Increased transport of pteridines compensates for mutations in the high affinity folate transporter and contributes to methotrexate resistance in the protozoan parasite *Leishmania tarentolae*.** *EMBO J.* 1999; **18**(9): 2342–51.
PubMed Abstract | Publisher Full Text | Free Full Text

123. Vasudevan G, Carter NS, Drew ME, *et al.*: **Cloning of *Leishmania* nucleoside transporter genes by rescue of a transport-deficient mutant.** *Proc Natl Acad Sci U S A.* 1998; **95**(17): 9873–8.
PubMed Abstract | Publisher Full Text | Free Full Text

124. Gazanion E, Fernandez-Prada C, Papadopoulou B, *et al.*: **Cos-Seq for high-throughput identification of drug target and resistance mechanisms in the protozoan parasite *Leishmania*.** *Proc Natl Acad Sci U S A.* 2016; **113**(21): E3012–21.
PubMed Abstract | Publisher Full Text | Free Full Text

125. F Tejera Nevado P, Bifeld E, Höhn K, *et al.*: **A Telomeric Cluster of Antimony Resistance Genes on Chromosome 34 of *Leishmania infantum*.** *Antimicrob Agents Chemother.* 2016; **60**(9): 5262–75.
PubMed Abstract | Publisher Full Text | Free Full Text | F1000 Recommendation

126. F De Rycker M, Hallyburton I, Thomas J, *et al.*: **Comparison of a high-throughput high-content intracellular *Leishmania donovani* assay with an axenic amastigote assay.** *Antimicrob Agents Chemother.* 2013; **57**(7): 2913–22.
PubMed Abstract | Publisher Full Text | Free Full Text | F1000 Recommendation

127. F Peña I, Pilar Manzano M, Cantizani J, *et al.*: **New compound sets identified from high throughput phenotypic screening against three kinetoplastid parasites: an open resource.** *Sci Rep.* 2015; **5**: 8771.
PubMed Abstract | Publisher Full Text | Free Full Text | F1000 Recommendation

128. F Khraiwesh M, Leed S, Roncal N, *et al.*: **Antileishmanial Activity of Compounds Derived from the Medicines for Malaria Venture Open Access Box Against Intracellular *Leishmania major* Amastigotes.** *Am J Trop Med Hyg.* 2016; **94**(2): 340–7.
PubMed Abstract | Publisher Full Text | Free Full Text | F1000 Recommendation

129. F Alsford S, Eckert S, Baker N, *et al.*: **High-throughput decoding of antitrypanosomal drug efficacy and resistance.** *Nature.* 2012; **482**(7384): 232–6.
PubMed Abstract | Publisher Full Text | Free Full Text | F1000 Recommendation

130. F Lye LF, Owens K, Shi H, *et al.*: **Retention and loss of RNA interference pathways in trypanosomatid protozoans.** *PLoS Pathog.* 2010; **6**(10): e1001161.
PubMed Abstract | Publisher Full Text | Free Full Text | F1000 Recommendation

131. Duncan SM, Myburgh E, Philipon C, *et al.*: **Conditional gene deletion with DiCre demonstrates an essential role for CRK3 in *Leishmania mexicana* cell cycle regulation.** *Mol Microbiol.* 2016; **100**(6): 931–44.
PubMed Abstract | Publisher Full Text | Free Full Text

132. F Kraeva N, Ishemgulova A, Lukeš J, *et al.*: **Tetracycline-inducible gene expression system in *Leishmania mexicana*.** *Mol Biochem Parasitol.* 2014; **198**(1): 11–3.
PubMed Abstract | Publisher Full Text | F1000 Recommendation

133. F Sander JD, Joung JK: **CRISPR-Cas systems for editing, regulating and targeting genomes.** *Nat Biotechnol.* 2014; **32**(4): 347–55.
PubMed Abstract | Publisher Full Text | Free Full Text | F1000 Recommendation

134. Selle K, Barrangou R: **Harnessing CRISPR-Cas systems for bacterial genome editing.** *Trends Microbiol.* 2015; **23**(4): 225–32.
PubMed Abstract | Publisher Full Text

135. Wright AV, Nuñez JK, Doudna JA: **Biology and Applications of CRISPR Systems: Harnessing Nature's Toolbox for Genome Engineering.** *Cell.* 2016; **164**(1–2): 29–44.
PubMed Abstract | Publisher Full Text

136. F Sollelis L, Ghorbal M, MacPherson CR, *et al.*: **First efficient CRISPR-Cas9-mediated genome editing in *Leishmania* parasites.** *Cell Microbiol.* 2015; **17**(10): 1405–12.
PubMed Abstract | Publisher Full Text | F1000 Recommendation

137. F Zhang WW, Matlashewski G: **CRISPR-Cas9-Mediated Genome Editing in *Leishmania donovani*.** *MBio.* 2015; **6**(4): e00861–15.
PubMed Abstract | Publisher Full Text | Free Full Text | F1000 Recommendation

138. F Peng D, Kurup SP, Yao PY, *et al.*: **CRISPR-Cas9-mediated single-gene and gene family disruption in *Trypanosoma cruzi*.** *MBio.* 2015; **6**(1): e02097–14.
PubMed Abstract | Publisher Full Text | Free Full Text | F1000 Recommendation

Open Peer Review

Current Referee Status: ✓ ✓ ✓ ✓ ✓

Editorial Note on the Review Process

F1000 Faculty Reviews are commissioned from members of the prestigious F1000 Faculty and are edited as a service to readers. In order to make these reviews as comprehensive and accessible as possible, the referees provide input before publication and only the final, revised version is published. The referees who approved the final version are listed with their names and affiliations but without their reports on earlier versions (any comments will already have been addressed in the published version).

The referees who approved this article are:

Version 1

1 **Gerald Spaeth**, Unité de Parasitologie moléculaire et Signalisation, Department of Parasites and Insect Vectors, Institut Pasteur, INSERM U1201, Paris, France
Competing Interests: No competing interests were disclosed.

2 **Yvon Sterkers**, UMR MIVEGEC (CNRS 5290 - IRD 224 - University Montpellier), University Regional Hospital of Montpellier, Montpellier, France
Competing Interests: No competing interests were disclosed.

3 **Richard McCulloch**, The Wellcome Trust Centre for Molecular Parasitology, Institute of Infection, Immunity and Inflammation, University of Glasgow, Glasgow, UK
Competing Interests: No competing interests were disclosed.

4 **Joachim Clos**, Bernhard Nocht Institute for Tropical Medicine, Hamburg, Germany
Competing Interests: No competing interests were disclosed.

5 **Francisco Gamarro**, Instituto de Parasitología Lopez-Neyra, IPBLN-CSIC, Granada, Spain
Competing Interests: No competing interests were disclosed.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 321291, 7 pages
http://dx.doi.org/10.1155/2015/321291

Review Article

The Genetics and the Genomics of Primary Congenital Glaucoma

Raffaella Cascella,[1,2] Claudia Strafella,[1] Chiara Germani,[3] Giuseppe Novelli,[1] Federico Ricci,[4] Stefania Zampatti,[3,5] and Emiliano Giardina[1,3]

[1] *Department of Biomedicine and Prevention, School of Medicine, University of Rome "Tor Vergata", 00133 Rome, Italy*
[2] *Emotest Laboratory, 80078 Pozzuoli, Italy*
[3] *Molecular Genetics Laboratory UILDM, Santa Lucia Foundation, 00142 Rome, Italy*
[4] *UOSD Retinal Pathology PTV Foundation "Policlinico Tor Vergata", 00133 Rome, Italy*
[5] *Neuromed IRCCS, 86077 Pozzilli, Italy*

Correspondence should be addressed to Raffaella Cascella; raffaellacascella@virgilio.it

Received 27 March 2015; Revised 17 July 2015; Accepted 12 August 2015

Academic Editor: Giacomina Massaro-Giordano

Copyright © 2015 Raffaella Cascella et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The sight is one of the five senses allowing an autonomous and high-quality life, so that alterations of any ocular component may result in several clinical phenotypes (from conjunctivitis to severe vision loss and irreversible blindness). Most parts of clinical phenotypes have been significantly associated with mutations in genes regulating the normal formation and maturation of the anterior segments of the eye. Among the eye anterior segment disorders, special attention is given to Glaucoma as it represents one of the major causes of bilateral blindness in the world, with an onset due to Mendelian or multifactorial genetic-causative traits. This review will point out the attention on the Primary Congenital Glaucoma (PCG), which is usually transmitted according to an autosomal-recessive inheritance pattern. Taking into consideration the genetic component of the PCG, it is possible to observe a strong heterogeneity concerning the disease-associated loci (*GLC3*), penetrance defects, and expressivity of the disease. Given the strong PGC heterogeneity, pre- and posttest genetic counseling plays an essential role in the achievement of an appropriate management of PCG, in terms of medical, social, and psychological impact of the disease.

1. Introduction

The sight is one of the five senses allowing an autonomous and high-quality life. Hence, sight defects may turn out to be really restricting for the life quality, representing a discriminating or even a life-threatening issue, especially in developing countries. As a matter of fact, about the 75% of information we receive about the world around us relies on the ability of the eyes to receive the light and transform it into a "picture" [1].

Given the critical importance of the sight organ, it is not surprising that alterations or malformations of any ocular component may result in several clinical phenotypes, ranging from conjunctivitis to Cataract until the severe vision loss and the irreversible blindness. Ocular clinical syndromes may develop owing to environmental agents (infections acquired during pregnancy or later in the life, aging processes, and systemic disorders), which can affect the ocular structures at the developmental or mature stages. However, most parts of clinical phenotypes have been significantly associated with mutations in genes regulating the normal formation and maturation of both posterior and anterior segments of the eye [2]. This review will be particularly focused on the genetics and genomics of the anterior segment disorders.

2. Genetic and Phenotypic Overview on the Eye Anterior Segment Mendelian Disorders

The clinical manifestations characterizing the anterior segment disorders reflect both phenotypic and genotypic heterogeneity, with multiple overlapping symptoms and associated genes [3]. Firstly, it is important to distinguish the anterior segment disorders characterized by a Mendelian or simple genetics from the disorders developed in consequence of complex or multifactorial traits. In particular, while the Mendelian diseases are due to causative rare mutations in specific genes (causative gene) and are inherited according to Mendel's laws, multifactorial diseases are caused by an interplay between several genetic and environmental triggering factors [2, 4]. In the large majority of cases, the genetic variants, associated with multifactorial diseases, are not rare mutations but rather polymorphisms, located on susceptibility genes and reported at high frequencies in normal/nonaffected population. These genetic markers are not sufficient to cause the disease but they only represent risk factors, determining a higher susceptibility to a particular disease. As mentioned below, it is interesting to notice that the presence of modifier genes, polymorphisms, or mutations may also be responsible for the variable expressivity occurring in Mendelian and complex diseases [5, 6]. In addition, the recognition of the distinctive features related to each anterior segment disorder can be complicated by different factors [5] such as

(i) the presence of extraocular (systemic) clinical manifestations, as facial dysmorphisms, dental dysplasia, redundant preumbilical skin, and short stature;

(ii) heterogeneous inheritance patterns (mainly autosomal-dominant and/or autosomal-recessive inheritance patterns);

(iii) developmental abnormalities occurring during embryogenesis and maturation of the eye anterior segment tissues (leading to congenital diseases);

(iv) foreign events, such as inflammations, infections, and trauma.

Among the most studied Mendelian diseases, it is important to mention Axenfeld-Rieger syndrome (ARS, #180500, #601499, and #602482), a rare disorder which affects 1 : 200.000 individuals and follows an autosomal-dominant inheritance pattern. The ARS etiopathology is characterized by an abnormal development of the anterior chamber, the cornea, and the iris. Clinical hallmarks of the disease include hypoplasia, corectopia, polycoria, posterior embryotoxon, and peripheral anterior synechiae (iris strands). Facial, dental, umbilical, and skeletal defects are also usually present as extraocular manifestations [5, 7].

Another Mendelian disorder affecting the anterior segment of the eye is referred to as Peters' Anomaly (#604229). It presents a heterogeneous inheritance, mostly autosomal-recessive, although autosomal-dominant and sporadic cases have been reported. The disease is caused by an incomplete separation of the cornea from the iris or the lenses, resulting in a variable central corneal opacity among patients. Systemic abnormalities reported in patients with Peters' Anomaly are short stature, cleft lip palate, growth, and mental delay [5, 8].

The complete or partial absence of the iris results in a pathological phenotype known as Aniridia (#106210). It affects 1 : 50.000–100.000 newborns worldwide and it is inherited in autosomal-dominant pattern. Aniridia is characterized by a reduced visual acuity and photophobia. Suddenly, it is associated with behavioral problems, developmental delay, and difficulties in detecting the odors [5, 9].

Concerning the disease phenotypes associated with malformations of the sclera and the cornea, it is important to mention the Sclerocornea and the Megalocornea. Sclerocornea (#18170) is a congenital disorder characterized by the absence of separation between sclera and cornea, which presents, as a consequence, a flat curvature. The malformation results in a nonprogressive corneal opacification (peripheral, sectoral, or central) [5, 10]. Megalocornea (#309300) instead is generally inherited in an X-linked recessive pattern and it consists of an enlargement of the corneal diameter (>13 mm) at birth, a deep anterior chamber, and normal IOP [5, 11].

Among the anterior segment disorders, particular attention has been paid to Keratoconus and Glaucoma, since both of the diseases are characterized by the presence of both Mendelian and multifactorial traits which give rise to different forms of Keratoconus and Glaucoma.

Keratoconus (#148300) is a noninflammatory degenerative disorder of the cornea affecting 1 in 2.000 individuals. The disease is due to the bulging and distortion of the corneal curvature and surface, which lead to the loss of visual acuity. A number of environmental factors have been associated with the development of Keratoconus, especially long-term contact lens wear, chronic eye rubbing, and atopy of the eye. However, the observation of familiarity transmission and studies on monozygotic and dizygotic twins have demonstrated that Keratoconus is characterized by a significant genetic component. In fact, the disease showed to be mostly transmitted in an autosomal-dominant inheritance pattern, although reduced penetrance and autosomal-recessive transmission have been reported [12].

Concerning Glaucoma several types of diseases can be distinguished, depending on the presence of causative Mendelian or multifactorial characters. In multifactorial Glaucoma, the clinical phenotype originates from the interaction between genetic/molecular (*LOXL1* genetic polymorphisms/retinal ganglionic cell death) and environmental (aging, diet, stress, and life-style) factors [13–15]. Although the hereditary component of the disease remains poorly understood, it is known that it can affect both genders, with higher risk for male subjects, advanced age people, and African descendants [16]. This form of Glaucoma is quite similar to several multifactorial, chronic, and age-related degenerative diseases, such as Osteoporosis, Osteoarthritis, Alzheimer's Disease, Coronary Artery Disease, Cataract [17]. In fact, multifactorial Glaucoma is a chronic neurodegenerative process, with an onset in the middle age and a slow progressive course, without particular signs or symptoms. Subjects affected by Glaucoma show a reduction of the visual field, blurry vision, decreased sensitivity to color, and contrast [16]. Over the multifactorial form, it is important to remark the existence of Glaucomas which are characterized by Mendelian inheritance (mainly autosomal-recessive and autosomal-dominant). Further details about clinical and genetic profile of Mendelian disease will be hereafter given.

Over the phenotypic manifestations of the anterior segment diseases, it is important to take into consideration the genetic background of these disorders. In fact, traditional linkage and positional cloning approaches identified a number of "modifier genes" associated with the development, inheritance pattern, variable expressivity, and phenotype of the anterior segment disorders. Such genetic modifiers consist of mutations or polymorphisms in specific genes, involved in the pathological mechanisms responsible for the disease phenotype [5, 13]. However, the identification of disease-causative genes led to the observation that not only is more than one gene mutation accountable for the same clinical condition, but also variable expressivity and incomplete penetrance can occur in patients carrying the same mutation. In addition, Genome Wide Association Studies (GWAS) allowed the identification of other genetic modifiers, especially common genetic variants associated with the individual susceptibility to eye anterior segment disorders. Each disease-associated individual variant confers a low risk when taken alone, while a set of low-risk variants provide a higher predictive value of the overall susceptibility to the disease [5]. The genes involved in the pathogenesis of these disorders include transcription factors in the majority of cases, although some genes code for transporters and glycosylation proteins. Given their function, mutations in the disease-causative genes will obviously have an impact on the downstream gene regulation, the integrity of the mature protein, and its final interaction with the cellular compartments [18]. The common low-risk variants instead are mostly located in noncoding regions (>80%) near the causative genes and essentially affect the overall gene expression patterns (regulation of RNA splicing and transcription activities) [5]. The variability of the gene expression profile can also be altered by epigenetic events (DNA methylation, histone acetylation/deacetylation, and structural chromatin modification), which act on the DNA transcription rather than on the primary DNA sequence [19]. In particular, DNA methyl transferases (DNMTs) were found to be differentially expressed in tissues and related cells of the eye anterior segment, such as the cornea, conjunctiva, anterior lenses, and trabecular meshwork [19]. Altered methylation profiles have been associated with few anterior segment disorders (such as Pterygium, #17800) [20]. Moreover, experiments on human cells and animal models demonstrated that histone modifications (HDA3) and miRNAs (miR29, miR204, miR146a, and miR24) play a role in the development of anterior segment diseases (as the Primary Open Angle Glaucoma, #137760) [21]. However, further research may help to study and discover epigenetic pathways concerning the etiopathogenesis of the congenital disorders affecting the eye anterior segment.

Independently of the phenotype/genotype features of the disorders, it is important to underline that a pathological phenotype does not result from the disruptive effect of one single gene mutation, but from the alteration of the network of genes to which the mutated gene specifically belongs. As a matter of fact, the high phenotypic and genotypic heterogeneity of the anterior segment disorders of the eye is actually the final outcome of the cross-talking between the disease-causative genes and the individual surrounding environment [5].

2.1. The Genes Causative of the Eye Anterior Segment Disorders. To date, a number of causative genes have been involved in the pathogenesis of the eye anterior segment disorders, some of which will be briefly described in the following paragraphs.

Both genes *PITX2* and *FOXC1*, located at 4q25 and at 6p25, respectively, code for two transcription factors (Pituitary Homeobox 2 and Forkhead Box C1). Linkage analysis has discovered that specific types of mutations in these genes are associated with specific pathological phenotype. In fact, intragenic mutations and/or deletions in *PITX2* and *FOXC1* have been reported in ARS syndrome with or without systemic anomalies. Copy Number Variation (CNV) has been found in ARS and Peters' Anomaly, while missense mutations have been identified in patients affected by Peters' Anomaly [3, 22].

PAX6 was the first gene to be associated with human anterior segment disorder. The gene is mapped on 11p13 and encodes for the transcription factor Paired Box 6. The mutational spectrum of the gene includes nonsense, splicing, insertion, and deletion mutations, resulting in the Aniridia disease. However, some missense mutations in *PAX6* have been reported in Peters' Anomaly and other ocular defects [18, 22].

B3GALTL is located on 13q12.3 and is responsible for the beta-1,3-glucosyltransferase (B3Glc-T) enzyme production, which is involved in the glycosylation process of proteins (posttranslational addition of sugar molecules). Deletion and splicing mutations have been reported in patients with Peter Plus Syndrome (PPS, #261540, Peters' Anomaly with variable systemic anomalies) [5, 22].

SOX2 is mapped on 3q26.3-q27 and codes for the transcription factor SRY-like box 2. It is especially involved in the development of the eyes, so that *SOX2* mutations have been associated with Sclerocornea and other ocular diseases [18].

CHRDL1 is located on Xq23 and encodes for the ventroptin protein, which works as bone morphogenic protein antagonist. Genomic techniques have associated mutation in *CHRDL1* with the X-linked form of Megalocornea [18].

VSX1 is situated on 20p11.21 and produces the transcription factor Visual System Homeobox 1. It is functionally involved in the regulation of the events occurring during craniofacial and ocular development. Missense mutations were found to be involved in the pathogenesis of Keratoconus, particularly concerning the autosomal-dominant form with variable expressivity and incomplete penetrance [12].

Other genes related to the above-mentioned and other eye anterior segment disorders (not described in this review) include *FOXE3*, *BMP4*, *BMP7*, *LAMB2*, *COL4A1*, *FGFR2*, *CYP1B1*, *LTBP2*, and *MYOC* [22].

In particular, this review will be focused on the genetic components underlining the congenital form of Glaucoma, a rare pathological condition affecting the anterior chamber drainage structures of the eye.

3. Focus on a Specific Disorder of the Anterior Segment of the Eye: The Glaucoma

As mentioned above, in the context of the eye anterior segment disorders special attention is given to Glaucoma. It represents one of the major causes of bilateral blindness in the world and several types of diseases have been classified to date, according to presence of Mendelian or multifactorial genetic-causative traits. In particular, Glaucoma consists of a pathological condition characterized by the progressive death of Retinal Ganglion Cells (RGCs), degeneration of the optical nerve, and vision loss as final result. The real explanation for RGCs death is not yet clear, although the level of IOP is known to play a fundamental role [13]. In particular, the IOP depends on the homeostasis between the aqueous humor produced in the ciliary body (and secreted in the posterior chamber) and its drainage through the trabecular meshwork in the anterior chamber angle. The balanced production and outflow of aqueous humor are constant and generate a positive pressure within the eye (~15 mmHg). Malfunctioning and/or malformations of the trabecular meshwork may result in the elevation of IOP over the limit ranges (>21 mmHg) and consequently the development of Glaucoma. Trabecular meshwork disruption can depend on a number of factors, such as mechanical and oxidative stress, aging, and genetic mutations [23]. The improper drainage of aqueous humor determines the increase of IOP, which in turn generates a mechanical stress and strain on the lamina cribrosa (the exit points of the optic nerve fibers in the sclera). The resulting stress blocks the transport of neurotrophic factors to the RGCs that finally die by apoptosis [24].

Glaucoma can be classified into primary, secondary, and primary congenital. The primary form is a nonsyndromic condition and it does not originate from any previous anterior segment abnormality, trauma, or inflammation. According to the compromised regions, two types of primary Glaucoma can be distinguished: Primary Closed-Angle Glaucoma and Primary Open-Angle Glaucoma. In the first condition, the drainage of aqueous humor is clogged by the closure of the angle between the iris and the cornea, while in the second form the fluid meets high resistance because of a malfunctioning of the trabecular meshwork [23].

The secondary Glaucoma instead is linked to the presence of ocular injuries or systemic conditions (for example, diabetes and long-term corticosteroid use) [23]. Even in this case, Closed-Angle and Open-Angle subtypes can be distinguished. On this subject, the Pseudoexfoliation Syndrome (PEXS) represents one of the major causes of secondary Open-Angle Glaucoma (accounting for the 25% of all Open-Angle Glaucomas). PEXS is described as an age-related systemic disease, which leads to the development of the secondary Glaucoma because of the accumulation of exfoliation material in the trabecular meshwork and consequently the increase of IOP [14].

This review will point out the attention on the Primary Congenital Glaucoma (PCG) which is a rare form representing the 1–5% of all cases of Glaucoma [25]. PCG generally presents an autosomal-recessive inheritance pattern, so that it affects only the 25% of newborns from parents carrying PCG-associated mutations. However, the incidence of PCG is highly heterogeneous in relation to the population, the geographic region, and the prevalence of consanguineous relationships. In fact, the incidence rate is estimated to be 1 : 2.500 in Saudi Arabia and Slovakia Gypsy populations (because of highly frequent consanguineous relations), in contrast with the lower incidence rate recorded among Western populations ranging from 1 : 18.500 and 1 : 30.000 [25, 26].

On the basis of the age of onset, it is possible to classify three subtypes of PCG: neonatal or newborn when presented at the birth or within the 1st month of life; infantile, diagnosed from 1st month to 2 years; and late-onset, recognized after 2 years. Usually, in 70–80% of cases, PCG affect both of the eyes (bilateral form) [27].

The clinical hallmarks of PCG diagnosis include IOP > 21 mmHg, optic cupping, epiphora, corneal edema and Haab striae, globe enlargement (buphthalmos), photophobia, blepharospasm, and hyperlacrimation. The PCG phenotype is due to a trabeculodysgenesis phenomenon, that is, an abnormal development of anterior chamber leading to the enlargement of the trabecular meshwork bundles and consequently reduction of the trabecular area available for the aqueous humor outflow [28]. In addition, the humor drainage is reduced because of the interference of immature iris, ciliary body, and anterior chamber angle structure which appear superimposed on the trabecular meshwork and compromise the aqueous humor outflow. The final result of these anterior segment structural defects and the altered aqueous humor drainage is the elevation of IOP and the enlargement of the entire ocular globe [25].

4. Genetics of PCG

Taking into consideration the genetic component of the PCG, it is possible to observe a strong heterogeneity in terms of disease-associated loci, penetrance defects, and expressivity of the disease among the different populations. Concerning the disease-associated loci, the Human Genome Organization established a specific nomenclature for Glaucoma-associated genetic loci. In fact, "GLC" stands for the general name of genes involved in Glaucoma; "1, 2, and 3" indicate the type of primary Glaucoma (Open-Angle, Closed-Angle, and congenital/infantile Glaucoma, resp.); "A, B, C, and D" refer to the progressive genes mapped for each Glaucoma type [26]. This review will focus particularly on *GLC3* loci, since these refer to the primary congenital Glaucoma.

4.1. GLC3 Loci. To date four PCG loci have been classified under *GLC3* subgroup, namely, *GLC3A*, *GLC3B*, *GLC3C*, and *GLC3D*.

The first locus to be associated with PCG was *GLC3A*, mapped on 2p21 chromosomal region. At this locus, about 147 mutations have been found in the gene *CYP1B1*, coding for the homonymous protein (cytochrome P450, family 1, subfamily B, and polypeptide 1) [29, 30]. *CYP1B1* consists of 3 exons and 2 introns, of which the first exon is a noncoding region, while the second and third exons are responsible for the production of CYP1B1 protein [25]. The *CYP1B1* mutations can be missense, nonsense, insertions, and deletions and result in the disruption of enzymatic activity and functionality of CYP1B1 protein [25, 31].

The mutations showed an autosomal-recessive inheritance pattern and are the most common variations identified among PCG patients, especially in consanguineous cohorts. However, a high variable distribution of *CYP1B1* mutations has been reported among the different worldwide populations, with 90–100% found in Saudi Arabia and Slovakia Gypsy populations, 14–30% among USA and European populations, and 15–20% reported in Japanese and Chinese populations [31–33]. Interestingly, in families carrying *CYP1B1* mutations cases of incomplete penetrance and variable expressivity have been observed. In fact, patients harboring the same mutations showed a different degree of disease severity, age of onset, or even lack of the disease phenotype [25, 26].

The pathogenetic role of the CYP1B1 protein in PCG has to be yet clarified, although a possible involvement in the metabolic pathways required for eye anterior chamber development is suggested, particularly for the trabecular meshwork formation [26]. On this subject, a recent study conducted on *Cyp1b1*$^{-/-}$ mice demonstrated that *Cyp1b1* deficiency was responsible for the increased oxidative stress and ultrastructural defects in the TM tissue of early-life animals. In particular, the absence of CYP1B1 protein (due to *Cyp1b1* silencing) hampers the proper removal of Reactive Oxygen Species (ROS) and, consequently, compromises the development and differentiation of TM tissue. In addition, the increased oxidative stress in *Cyp1b1*$^{-/-}$ mice caused the decrease of the Periostin (Postn) production, which results in the loss of the mechanical strength and structural integrity of the TM tissue [34].

GLC3B and *GLC3C* loci have been associated with the PCG, although no gene has been identified in both regions to date. In particular, *GLC3B* is located at 1p36.2–1p36.1 while *GLC3C* maps on 14q24.3–14q31.1 [25, 26].

Concerning the *GLC3D* locus (14q24), autosomal-recessive null mutations have been reported at this position and located in the *LTBP2* gene. *LTBP2* codes for the Latent Transforming beta Binding Protein 2, a matrix protein playing a role in tissue repair processes and cell adhesion. The role of the PCG still remains to be explained. However, it has been demonstrated that *LTBP2* is expressed in the trabecular meshwork and ciliary body and PCG-causative null mutations have been found in consanguineous Pakistani, European Gypsy, and Iranian families [26, 35].

4.2. Other PCG-Associated Genes. Families affected by PCG carried mutations in *MYOC* gene, independently of the presence or absence of *CYP1B1* mutations. *MYOC* is a 3-exon gene, located at 1q24.3–1q25.2 chromosomic region, and codes for the glycoprotein myocilin. The function of the protein is unknown, while its expression has been reported at high level in the trabecular meshwork and the ciliary body. Mutations in *MYOC* have been found to be a leading cause of aqueous outflow obstruction through the trabecular meshwork and of the consequent IOP elevation. A number of *MYOC* mutations have been associated with the juvenile- and adult-onset primary Glaucoma, characterized by an autosomal-dominant/autosomal-recessive inheritance pattern with high penetrance.

Concerning the role of *MYOC* in PGC, possible interactions with *CYP1B1* and/or other unknown loci have been hypothesized to work as genetic modifiers to determine an earlier onset of the disease in patients harboring mutations in *MYOC* and *CYP1B1*, with respect to the patients negative for *MYOC* or *CYP1B1* mutations [25, 26, 36].

Another putative PCG-associated gene is *FOXC1*, a protein expressed in periocular mesenchyme cells, producing ocular drainage structures as iris, cornea, and trabecular meshwork. Interestingly, a deletion in *FOXC1* has been detected in PCG and other ocular and nonocular defects, highlighting a possible involvement of *FOXC1* in the pathogenetic pathway of the disease [25, 26].

Mutations in *BMP4* gene (14q22-q23) have been identified in patients affected by PCG and other disease phenotypes. *BMP4* codes for the bone morphogenetic protein 4 and it is expressed in different tissues, among which is the ocular vesicle, and in the optic cup. However, further research is needed to clarify the role of *BMP4* in PCG pathogenesis [26].

5. The Genetic Counseling in the Management of PCG

As PCG is a severe disease characterized by a particular genetic base and familiarity, genetic counseling represents a useful tool for the disease management (clinical features of the disease, prevision of the disease progression, and medical complications) [37]. Genetic counseling is a communication activity oriented to help the patient and/or their family in receiving medical information concerning the genetic features associated with PCG. In particular, the genetic specialist has to explain the recurrence risk of PCG in relation to the presence of positive/negative familiarity and to the frequency of PCG in the general population. The construction of familiar pedigree may be of help for clarification of the genetic model of inheritance of the disease. In fact, although the majority of cases of PCG appear to be sporadic cases or to follow an autosomal-recessive inheritance pattern, in rare cases some families reported an autosomal-dominant inheritance model. After the clarification of the genetic base of the disease, the recurrence risk and familiarity for PCG have been assessed, and the specialist should suggest and give details about the possibility of performing a genetic test for the detection of PCG-causative mutations. In particular, as illustrated before, mutations in *CYP1B1*, *LTBP2*, and *MYOC* have been found to be causative of PCG among different worldwide populations [25]. However, the specialist has to put in evidence the variable expressivity and penetrance defects occurring in presence of causative mutations. In fact, it is important to clarify to the patient that the positivity for PCG-associated mutations does not always correspond to the development of the effective pathological phenotype and the disease may manifest at later age. In case of presence of PCG-associated mutations, the genetic specialist should suggest to the patient a constant follow-up of his own clinical conditions. This aspect is strongly important in case of siblings of children positive to *CYP1B1* mutations, where there is a 25% higher risk of PCG development because of the variable expressivity of the disease phenotype [25, 38].

To help understand the pros and cons of genetic tests, it is very important to perform the informative counseling prior and after the test (pre- and a posttest). In this context the explanation of biological aspects and technical approaches related to the genetic test may help in understanding limits of negative results, false-negatives, and the need of further investigations.

Unfortunately, some cases may result negative to the analysis of PCG-causative known mutations, so that undiscovered mutations in other unidentified genes may be possible (e.g., unknown genes mapped on *GLC3B* and *GLC3C* loci). Social and psychological implications of the lack of a responsible gene involve the impossibility to perform genetic tests for the prediction of the disease risk in families with a positive-PCG history. This aspect may appear as a "sword of Damocles" and may affect deeply the couple perspective in matter of family construction.

6. Conclusion

The visual loss is due to interruption/aberration/malformation of the normal functioning of the ocular structures. Early-onset ocular diseases usually have Mendelian inheritance, while common adult-onset disorders are inherited as complex traits. To date, several genetic and molecular studies have provided insights into the biological processes underlying many ophthalmic disorders. In this context, GWAS identified different genetic variants contributing to a number of common ocular complex disorders.

Glaucoma is a heterogeneous group of disorders causing irreversible blindness worldwide. As previously described, most forms of Glaucoma present a significant genetic component, characterized by the different causative or susceptibility genes identified through traditional (genetic linkage) or novel (GWAS) approaches. Among the different forms of Glaucoma, we focused our attention on the PCG, that is, the single most common childhood Glaucoma.

PCG is a rare disease affecting the structures of the anterior chamber of the eye anterior segment. Although rare, it is still the most common Glaucoma in infancy and causes a disproportionately high percentage of childhood blindness worldwide. It displays a strong heterogeneity in terms of disease phenotype and genotype. Genetically speaking, PCG presents a familiarity trend, with higher frequency in populations where consanguineous relationships are common (Saudi Arabia and Slovakia Gypsy populations). To date, several mutations in *CYP1B1*, *LTBP2*, and *MYOC* have been associated with the development of the disease, although the expressivity phenotype and the penetrance are variable. Given these data, pre- and posttest genetic counseling has an essential role for an adequate management of PCG, in terms of medical, social, and psychological impact of the disease.

It is important to remark that the eye has been at the forefront of translational gene therapy largely because of the availability of appropriate disease targets and its suitable anatomic features. These advantages have further fostered the research that culminated in the establishment of various clinical trials for the gene therapy of ocular diseases. In this perspective the identification of modifier genes should be encouraged as well as the analysis of the interaction between genes and environmental factors. The challenge for the future is to classify patients in relation to their genetic background, in order to provide personal treatment and management.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] J. Graw, "The genetic and molecular basis of congenital eye defects," *Nature Reviews Genetics*, vol. 4, no. 11, pp. 876–888, 2003.

[2] R. Cascella, M. Ragazzo, C. Strafella et al., "Age-related macular degeneration: insights into inflammatory genes," *Journal of Ophthalmology*, vol. 2014, Article ID 582842, 9 pages, 2014.

[3] J. C. Sowden, "Molecular and developmental mechanisms of anterior segment dysgenesis," *Eye*, vol. 21, no. 10, pp. 1310–1318, 2007.

[4] E. Giardina, C. Sinibaldi, and G. Novelli, "The psoriasis genetics as a model of complex disease," *Current Drug Targets: Inflammation and Allergy*, vol. 3, no. 2, pp. 129–136, 2004.

[5] Y. A. Ito and M. A. Walter, "Genomics and anterior segment dysgenesis: a review," *Clinical and Experimental Ophthalmology*, vol. 42, no. 1, pp. 13–24, 2014.

[6] E. Giardina, C. Sinibaldi, and G. Novelli, "Mapping the future of common diseases: lessons from psoriasis," *Frontiers in Bioscience*, vol. 12, no. 4, pp. 1563–1573, 2007.

[7] M. B. Shields, E. Buckley, G. K. Klintworth, and R. Thresher, "Axenfeld-Rieger syndrome. A spectrum of developmental disorders," *Survey of Ophthalmology*, vol. 29, no. 6, pp. 387–409, 1985.

[8] C. Kupfer, T. Kuwabara, and W. J. Stark, "The histopathology of Peters' anomaly," *American Journal of Ophthalmology*, vol. 80, no. 4, pp. 653–660, 1975.

[9] H. Lee, R. Khan, and M. O'keefe, "Aniridia: current pathology and management," *Acta Ophthalmologica*, vol. 86, no. 7, pp. 708–715, 2008.

[10] J. H. Elliot, S. S. Feman, D. M. O'Day, and M. Garber, "Hereditary sclerocornea," *Archives of Ophthalmology*, vol. 103, no. 5, pp. 676–679, 1985.

[11] D. A. Mackey, R. G. Buttery, G. M. Wise, and M. J. Denton, "Description of X-linked megalocornea with identification of the gene locus," *Archives of Ophthalmology*, vol. 109, no. 6, pp. 829–833, 1991.

[12] K. K. Abu-Amero, A. M. Al-Muammar, and A. A. Kondkar, "Genetics of keratoconus: where do we stand?" *Journal of Ophthalmology*, vol. 2014, Article ID 641708, 11 pages, 2014.

[13] D. B. Gould and S. W. M. John, "Anterior segment dysgenesis and the developmental glaucomas are complex traits," *Human Molecular Genetics*, vol. 11, no. 10, pp. 1185–1193, 2002.

[14] E. Giardina, F. Oddone, T. Lepre et al., "Common sequence variants in the *LOXL*1 gene in pigment dispersion syndrome and pigmentary glaucoma," *BMC Ophthalmology*, vol. 14, no. 52, pp. 1–6, 2014.

[15] J. L. Wiggs, "The cell and molecular biology of complex forms of glaucoma: updates on genetic, environmental, and epigenetic risk factors," *Investigative Ophthalmology & Visual Science*, vol. 53, no. 5, pp. 2467–2469, 2012.

[16] P. Azarbod, L. Crawley, F. Ahmed, M. F. Cordeiro, and P. Bloom, "Recent advances in the diagnosis and management of glaucoma," *Prescriber*, vol. 26, no. 1-2, pp. 21–25, 2015.

[17] X.-H. Xu, S.-S. Dong, Y. Guo et al., "Molecular genetic studies of gene identification for osteoporosis: the 2009 update," *Endocrine Reviews*, vol. 31, no. 4, pp. 447–505, 2010.

[18] E. James, H. Smith, and I. Elias, "Malformation of the anterior segment of the eye," in *Genetic Disease of the Eye*, E. I. Trabulosi, Ed., chapter 7, pp. 92–108, Oxford University Press, 2011.

[19] N. Bonnin, C. Belville, F. Chiambaretta, V. Sapin, and L. Blanchon, "DNA methyl transferases are differentially expressed in the human anterior eye segment," *Acta Ophthalmologica*, vol. 92, no. 5, pp. e366–e371, 2014.

[20] M. M. Liu, C.-C. Chan, and J. Tuo, "Epigenetics in ocular diseases," *Current Genomics*, vol. 14, no. 3, pp. 166–172, 2013.

[21] B. Yan, J. Yao, Z.-F. Tao, and Q. Jiang, "Epigenetics and ocular diseases: from basic biology to clinical study," *Journal of Cellular Physiology*, vol. 229, no. 7, pp. 825–833, 2014.

[22] L. M. Reis and E. V. Semina, "Genetics of anterior segment dysgenesis disorders," *Current Opinion in Ophthalmology*, vol. 22, no. 5, pp. 314–324, 2011.

[23] Y. A. Ito and M. A. Walter, "Genetics and environmental stress factor contributions to anterior segment malformations and glaucoma," in *Glaucoma—Basic and Clinical Aspects*, S. Rumelt, Ed., chapter 3, pp. 27–56, InTech, Rijeka, Croatia, 2013.

[24] R. N. Weinreb, T. Aung, and F. A. Medeiros, "The pathophysiology and treatment of glaucoma: a review," *The Journal of the American Medical Association*, vol. 311, no. 18, pp. 1901–1911, 2014.

[25] R. Sharafieh, A. H. Child, and M. Sarfarazi, "Molucular genetics of primary congenital glaucoma," in *Genetic Disease of the Eye*, Trabulosi, Ed., chapter 17, pp. 295–307, 2011.

[26] A. O. Khan, "Genetics of primary glaucoma," *Current Opinion in Ophthalmology*, vol. 22, no. 5, pp. 347–355, 2011.

[27] S. L. Zagora, C. L. Funnell, F. J. Martin et al., "Primary congenital glaucoma outcomes: lessons from 23 years of follow-up," *American Journal of Ophthalmology*, vol. 159, no. 4, pp. 788.e2–796.e2, 2015.

[28] A. K. Mandal and D. Chakrabarti, "Update on congenital glaucoma," *Indian Journal of Ophthalmology*, vol. 59, pp. S148–S157, 2011.

[29] D. B. Moore, O. Tomkins, and I. Ben-Zion, "A review of primary congenital glaucoma in the developing world," *Survey of Ophthalmology*, vol. 58, no. 3, pp. 278–285, 2013.

[30] I. Stoilov, A. N. Akarsu, and M. Sarfarazi, "Identification of three different truncating mutations in cytochrome P4501B1 (CYP1B1) as the principal cause of primary congenital glaucoma (Buphthalmos) in families linked to the GLC3A locus on chromosome 2p21," *Human Molecular Genetics*, vol. 6, no. 4, pp. 641–647, 1997.

[31] S.-H. Lim, K.-N. Tran-Viet, T. L. Yanovitch et al., "*CYP1B1*, *MYOC*, and *LTBP2* mutations in primary congenital glaucoma patients in the United States," *American Journal of Ophthalmology*, vol. 155, no. 3, pp. 508–517, 2013.

[32] K. K. Abu-Amero, E. A. Osman, A. Mousa et al., "Screening of CYP1B1 and LTBP2 genes in Saudi families with primary congenital glaucoma: genotype-phenotype correlation," *Molecular Vision*, vol. 17, pp. 2911–2919, 2011.

[33] J. L. Wiggs, A. M. Langgurth, and K. F. Allen, "Carrier frequency of *cyp1b1* mutations in the United States (an American ophthalmological society thesis)," *Transactions of the American Ophthalmological Society*, vol. 112, pp. 94–102, 2014.

[34] Y. Zhao, C. Sorenson, and N. Sheibani, "Cytochrome P450 1B1 and primary congenital glaucoma," *Journal of Ophthalmic & Vision Research*, vol. 10, no. 1, pp. 60–67, 2015.

[35] M. Ali, M. McKibbin, A. Booth et al., "Null mutations in LTBP2 cause primary congenital glaucoma," *The American Journal of Human Genetics*, vol. 84, no. 5, pp. 664–671, 2009.

[36] K. Kaur, A. B. M. Reddy, A. Mukhopadhyay et al., "Myocilin gene implicated in primary congenital glaucoma," *Clinical Genetics*, vol. 67, no. 4, pp. 335–340, 2005.

[37] A. Ganesh and A. Al-Mujaini, "Ocular genetics: a sub-specialty service for genetic eye diseases," *Oman Medical Journal*, vol. 28, no. 1, pp. 1–2, 2013.

[38] N. Tamçelik, E. Atalay, S. Bolukbasi, O. Çapar, and A. Ozkok, "Demographic features of subjects with congenital glaucoma," *Indian Journal of Ophthalmology*, vol. 62, no. 5, pp. 565–569, 2014.

frontiers
in Oncology

REVIEW
published: 06 July 2021
doi: 10.3389/fonc.2021.699265

Imaging-Genomics in Glioblastoma: Combining Molecular and Imaging Signatures

Dongming Liu[1†], *Jiu Chen*[2,3†], *Xinhua Hu*[1,3], *Kun Yang*[1], *Yong Liu*[1], *Guanjie Hu*[1], *Honglin Ge*[1], *Wenbin Zhang*[1,3*] and *Hongyi Liu*[1,3*]

[1] Department of Neurosurgery, The Affiliated Brain Hospital of Nanjing Medical University, Nanjing, China, [2] Institute of Neuropsychiatry, The Affiliated Brain Hospital of Nanjing Medical University, Fourth Clinical College of Nanjing Medical University, Nanjing, China, [3] Department of Neurosurgery, Institute of Brain Sciences, The Affilated Nanjing Brain Hosptial of Nanjing Medical University, Nanjing, China

Based on artificial intelligence (AI), computer-assisted medical diagnosis can scientifically and efficiently deal with a large quantity of medical imaging data. AI technologies including deep learning have shown remarkable progress across medical image recognition and genome analysis. Imaging-genomics attempts to explore the associations between potential gene expression patterns and specific imaging phenotypes. These associations provide potential cellular pathophysiology information, allowing sampling of the lesion habitat with high spatial resolution. Glioblastoma (GB) poses spatial and temporal heterogeneous characteristics, challenging to current precise diagnosis and treatments for the disease. Imaging-genomics provides a powerful tool for non-invasive global assessment of GB and its response to treatment. Imaging-genomics also has the potential to advance our understanding of underlying cancer biology, gene alterations, and corresponding biological processes. This article reviews the recent progress in the utilization of the imaging-genomics analysis in GB patients, focusing on its implications and prospects in individualized diagnosis and management.

Keywords: glioblastoma, radiomics, imaging genomics, artificial intelligence, machine learning, deep learning

OPEN ACCESS

Edited by:
Nico Sollmann,
University of California, San Francisco,
United States

Reviewed by:
Milan Girish Chheda,
Washington University School of
Medicine in St. Louis, United States
Manjari Pandey,
West Cancer Center, United States
Vasileios Stathias,
University of Miami, United States

***Correspondence:**
Hongyi Liu
njnkyylhy@163.com
Wenbin Zhang
wenbinzhang@njmu.edu.cn

[†]*These authors have contributed equally to this work and share first authorship*

Specialty section:
This article was submitted to
Neuro-Oncology and
Neurosurgical Oncology,
a section of the journal
Frontiers in Oncology

Received: *23 April 2021*
Accepted: *23 June 2021*
Published: *06 July 2021*

Citation:
Liu D, Chen J, Hu X, Yang K, Liu Y, Hu G, Ge H, Zhang W and Liu H (2021) Imaging-Genomics in Glioblastoma: Combining Molecular and Imaging Signatures. Front. Oncol. 11:699265. doi: 10.3389/fonc.2021.699265

Abbreviation: AI, Artificial intelligence; AUC, Area under receiver operating characteristic curve; CNNs, Convolutional neural networks; CNS, Central nervous system; CpG, Cytosine-phosphate-guanine; EGFR, Epidermal growth factor receptor; FLAIR, Fluid-attenuated inversion recovery; GB, Glioblastomas; GO, Gene Ontology; GSEA, Gene set enrichment analysis; 2HG, R(-)-2-hydroxyglutarate; IBSI, Image biomarker standardization initiative; IDH, Isocitrate dehydrogenase; LASSO, Least absolute shrinkage and selection operator; MRI, Magnetic resonance imaging; MGMT, O(6)-methylguanine-DNA methyltransferase; NF1, Neurofibromatosis 1; OS, Overall survival; PET, Positron emission tomography; PFS, Progression-free survival; PTEN, Gene of phosphate and tension homology deleted on chromosome ten; PIK3, Phosphatidylinositol-3-OH kinase; RB1, Retinoblastoma 1; RF, Random forest; RNN, Recurrent neural network; ROI, Region of interest; RRS, Radiomic risk score; RTK, Receptor tyrosine kinase; SMOTE, Synthetic minority oversampling technique, SVM, Support vector machines; TCGA, The Cancer Genome Atlas; TCIA, The Cancer Imaging Archive; TMZ, Temozolomide; TPOT, Tree-based Pipeline Optimization Tool; WHO, World Health Organization.

INTRODUCTION

Glioblastoma (GB) is the most common malignant primary brain tumor in adults, accounts for 48.3% of malignant brain tumors (1). Much work has focused on GB, given that it has the characteristics of infiltrating the surrounding brain tissue (2), and has an aggressive course and extremely grim prognosis (3). Gene expression profiling has provided new perspectives for the identification of tumor-associated genetic mutation, biomarkers, and therapeutic targets of glioblastoma (4, 5). The 2016 central nervous system (CNS) tumor classification of the World Health Organization (WHO) incorporated molecular markers into traditionally histopathological classification, providing an integrated phenotypic and genetic classification (6). Recently, based on transcriptomic profiling, GBs have been divided into proneural, classical, and mesenchymal subtypes (7).

The characteristic neuroradiological findings raise the initial suspicion for GB and determine the initial surgical approach and plan. Diagnosis is confirmed through histo-molecular evaluation of the tissue post-biopsy/resection. Traditional magnetic resonance imaging (MRI) has greatly improved the diagnostic efficiency of intracranial lesions, but there are still limitations in the definitive diagnosis of brain tumors. Over the past few decades, there has been a rapid development towards the application of quantitative parametric MRI techniques for GB evaluation (8). More recently, advanced computer technology has permeated the field of modern medicine, making precision medicine an inevitable trend in the development of clinical medicine. Artificial intelligence (AI) is a broad term that covers a wide range of disciplines including computer vision, machine learning, and more. Deep learning is a branch of machine learning that uses artificial neural networks as a framework to learn and characterize databases. There has been growing interest in the application of AI techniques to neuroimaging research, making machine learning and deep learning algorithms critical to the radiomics procedures.

Radiomics utilizes high-throughput radiomics features and mathematical models to quantify tumor characteristics, allowing the non-invasive capture of microscale information hidden within medical imaging (9, 10). As an important branch of radiomics, imaging-genomics (also known as radiogenomics) further links imaging characteristics (phenotypes) with genetic, mutational, and expression patterns (11, 12). The basic hypothesis of imaging-genomics is that the expression of a specific set of genes or molecular alterations affects the extractable imaging phenotypes (13). A particular focus of imaging-genomics analysis has been on the association between imaging characteristics and gene expression patterns, including the expression of individual genes and the expression of specific gene subgroups (14). Just like the resolution and complexity that genomics has brought to tumor biology, imaging-genomics also brings similar effects to traditional medical images. The ultimate goal of radiomics is to maximize the utilization of medical imaging information to help clinicians make clinical decisions, so as to minimize the possibility of potential invasive maneuvers. The increasing importance of genetic markers has also led to the rapid growth of GB imaging-genomics research, with progressively greater complexity, including the combination of artificial intelligence (AI) technology. More recently, deep learning using convolutional neural networks has further improved radiogenomics prediction (15, 16), providing unique opportunities and challenges for the diagnosis and treatment evaluation of GB.

Herein, we will review the latest researches on radiomics and imaging-genomics of GB. First, we briefly introduced the genomic characteristics of GB and the principle of imaging-genomics, and then summarized the findings of the latest research. Finally, we will present our opinions on its current challenges, and potential future directions.

GENOMICS IN GLIOBLASTOMA

Traditionally, GBs can be divided into "primary" and "secondary" GBs. The vast majority of GB (about 90%) develop rapidly *de novo* in elderly patients and are mostly isocitrate dehydrogenase (IDH) wild-type. While secondary GBs develops from a preexisting lower-grade glioma in younger patients and usually carry mutations in IDH (17). Radiologically, primary GBs are widely distributed in the brain, while secondary GBs are preferentially located in the area of the frontal lobe surrounding the rostral extension of the lateral ventricle in younger patients (17, 18). Primary and secondary GBs are largely histologically indistinguishable, but they are different in genetic and epigenetic profiles (17). The genomic and transcriptome characteristics of these tumors have revealed the key alterations that may contribute to the classification and evaluation of the disease (4, 19, 20) (17) (21).Different bioinformatics algorithms have identified several genes as significantly mutated in GB, including but not limited to epidermal growth factor receptor (EGFR), TP53, PTEN, Neurofibromatosis 1 (NF1), and RB1 (20, 21). Additionally, genomic analysis based on multiple sampling from the same neoplasm suggests that transcriptome subtypes are heterogeneous in GB (22) (23). Patient-specific intratumor heterogeneity of GB underlies variation in response to treatment, which contributes to the eventual failure of treatment (22, 24, 25), including drug resistance, radiotherapy resistance, and fast and incurable tumor recurrence. Thus, relying only on single tissue samples from individual patients to understand the cancer dynamics of GB is challenging.

The whole-genome sequence analysis has revealed the genetic and epigenetic landscapes of human brain GB (4, 20). Complex gene interaction events and molecular modulation networks are involved in the occurrence and development of GBs (4, 5, 26). It is expected that one or several biomarkers will provide information to help clinicians make the diagnosis, and enable clinical experts to make the best choice among various treatment options, such as surgery and/or chemotherapy, and/or radiotherapy, and/or targeted therapy. Numerous studies have identified alterations of several core signaling pathways in GB, including the RB1 pathway, the TP53 pathway, and the PI3K/PTEN pathway (4, 21). The heterogeneity of GB cells (26) suggests that targeting strategies based on multiple elements of different signaling pathways might constitute a more efficient therapy for GB (5).

Given its invasive and infiltrating nature, GB is essentially considered to be a disease of the entire brain (27). Surgical biopsy is currently the standard procedure for providing the genomics information of GB. However, a biopsy is not only highly invasive, but its accuracy is also limited by the biopsy site, resection type, and disease variability. Compared with traditional genomic analysis, imaging-genomics has the advantage of assessing the entire tumor volume, and non-invasively provides a "virtual biopsy" by predicting specific gene status, biological processes, and even core signaling pathways of GB (28–30) (28, 29, 31). High-resolution imaging technologies can present the tumor completely and three-dimensionally. Based on this property, imaging-genomics can associate the radiological phenotypes (such as shape or texture) of a specific spatial location (such as necrosis, high perfusion, or even more subtle subregions) with the tumoral molecular characteristics of the same location. Also, by associating the genetic status of the biopsy location with the quantitative imaging features, imaging-genomics has the potential to help characterize tumoral genetic heterogeneity (30), which shows diagnostic value under the context of individualized oncology.

RADIOMICS AND IMAGING-GENOMICS OF GLIOBLASTOMA

Traditional pathology and radiology have been primarily focused on the associations between histopathologic and visual imaging findings, while radiomics and imaging-genomics focus on exploring the relationships between imaging phenotypes and biological characteristics (11, 12). Generally, radiomics is dedicated to discovering the biological significance (such as predicting survival and treatment prognosis) of specific imaging phenotype, which when correlated with genomics is termed imaging-genomics. Imaging-genomics analysis helps to understand the biological associations behind image phenotypes, explain how biological processes are reflected in imaging, and define imaging markers associated with molecular biological characteristics. When constructing an imaging genomic map of GB, several dimensions need to be considered, including the disease course, imaging category, lesion delineation, features extraction and selection, and biological data type, as well as model building or integration (11). Typically, the imaging features used in a radiomics or imaging-genomics study are mainly divided into two categories, namely manually defined features and deep learning features.

Feature Extraction, Selection, and Corresponding Modeling

Manually Defined Features

Manually defined feature-based imaging-genomics utilizes specific mathematical algorithms to extract features from pre-depicted regions of interest (ROI) (**Figure 1**). After feature extraction and feature selection, machine learning models are used to solve classification and/or regression problems. Usually, the steps of the process mentioned above includes a few phases as following: 1) Data acquisition; 2) Image pre-processing, including multiparametric imaging registration, noise reduction, intensity and/or orientation normalization, spatial resampling, and corrections of MRI field inhomogeneities; 3) Segmentation of ROI, segment the tumors into necrosis, enhancement, edema, and peritumoral parts, etc. by manual

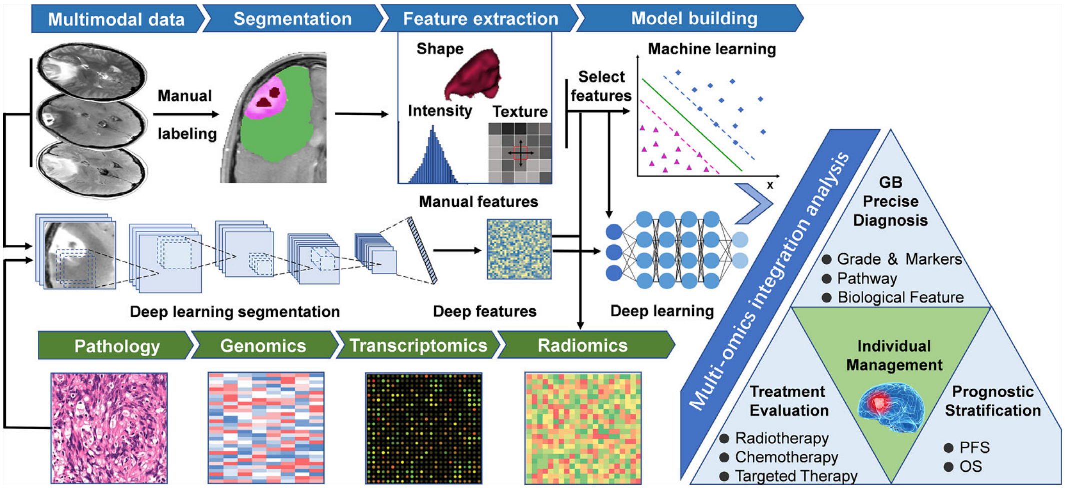

Multimodal data
Segmentation
Feature extraction
Model building
Manual
labeling
Shape
Intensity
Texture
Select
features
Machine learning
x
Manual features
Deep learning segmentation
Deep features
Deep learning
Pathology
Genomics
Transcriptomics
Radiomics
Multi-omics integration analysis
GB
Precise
Diagnosis
● Grade & Markers
● Pathway
● Biological Feature
Individual
Management
Treatment
Evaluation
● Radiotherapy
● Chemotherapy
● Targeted Therapy
Prognostic
Stratification
● PFS
● OS

FIGURE 1 | General workflow of radiomics studies in neuro-oncology. The workflow of a radiomics study, including the following steps: (1) Multimodal imaging and biological data acquisition; (2) Data preprocessing and standardization; (3) Delineation of regions of interest, including manual segmentation and deep learning-based segmentation; (4) Radiomics feature extraction using predefined algorithms or deep learning techniques; (5) Data analysis, feature reduction, and/or selection for further analysis of machine learning and/or deep neural networks; (6) Multi-omics and clinical information integrated model training and testing, guiding individualized disease diagnosis, treatment evaluation, and prognosis prediction. GB, Glioblastoma; OS, Overall Survival; PFS, Progression free survival.

delineation or (semi-) automated algorithms; 4) Feature extraction, extract different kinds of quantitative features that reflect the heterogeneity of tumors, including shape features, histogram-based features, textural features, and higher-order statistics features; 5) Feature selection, including supervised and unsupervised methods to overcome the issue of redundancy and overfitting that may be seen with multitudinous features; 6) Model building and evaluation, build a mathematical model (such as regression models, random forest (RF), and/or support vector machines (SVM) based on the selected features to predict a known, underlying ground truth, including specific genetic mutation or pathway alterations. In general, training set data is used for learning to fit the parameters of a classifier, while the validation set is used to tune the parameters of the model. Ideally, after model training and validation, the final models should be applied to the test set (unexposed independent data) to evaluate the performance of the final model. A robust radiomic model must be validated by independent, external data to show its value for clinical promotion and application.

Deep Learning-Based Features

Deep learning, a subcategory of machine learning based on neural networks (32), represents a class of algorithms that use stacked neural network structures (33). In deep learning-based imaging-genomics, different network structures such as auto-encoders or convolutional neural networks (CNNs) are utilized to mine the hidden features from the initial data (15, 34). The typical way to extract deep learning features is to train a deep learning model using observed indicators, and then extract the outputs of specific layers as deep learning features (35). For example, the results of the final fully connected layer of CNNs or the final convolutional layer of auto-encoder can be extracted as deep learning features (36). The extracted deep features can either be trained by an artificial neural network for classification and prediction (15) or leave the network and fit into different classifiers (35) (such as SVM, RF) to generate a model (**Figure 1**), the latter process is similar to the approach of manually defined feature-based imaging-genomics. Recently, deep learning has been successfully applied in digital pathology to segment nuclei and for the classification of molecular markers (37, 38). Deep learning-based imaging-genomics usually requires relatively large datasets. Auto-encoders provide an alternative method to extract deep features when there is not enough data to train a deep learning model (34, 36). Another technique to overcome this limitation is transfer learning (39, 40), which essentially uses a pre-trained network (usually on natural images) to solve the data set needs of deep network training (34). **Figure 1** describes the main workflow of the above two methods, and **Table 1** summarizes several radiomics studies designed according to this workflow.

Imaging-Genomics of Glioblastoma

As a subclass of radiomics, imaging-genomics focuses more on the relationships between imaging characteristics of a disease, and its gene expression patterns, gene mutations, and other genomic-related characteristics (11). The gene expression patterns for constructing these relationships include not only the expression of a single gene (45, 50), but also the measures that summarize expressions of specific gene subsets (14, 29, 51, 52), the corresponding signal pathways (29, 53), and consequent biological effects (28, 54). 'Imaging-genomics' research also refers to a study aimed at discovering those relationships.

In addition to predicting the expression status of specific molecular markers, the current imaging genomics researches can

TABLE 1 | Applications of radiomics for predicting specific molecular markers in GB.

Study	Imaging Modality	Molecular signature(s)	No. of Patients Training + Testing	No. of features Initial + Final		Performance AUC/Accuracy	Features selection and Model building
Li et al. (41)	MRI, T1+T1CE+T2+Flair	IDH1 status	225 (118 + 107)	1614	8	0.96^{a}	RF-Boruta, RF
Chang et al. (42)	MRI, T1+T1CE+T2+Flair	IDH1,MGMT,and 1p/19q	259 (80%+20%)	NA	NA	(IDH1) 94%^{b}	CNNs
Zhang et al. (43)	MRI,T1+T1CE+T2+ Flair +DWI	IDH1/2	120 (90 + 30)	2970	387	0.92^{a}	Correlation coefficient, RF
Li et al. (44)	PET	IDH1/2	127 (84 + 43)	1561	11	0.90^{a}	Lasso, Multivariate logistic regression
Wei et al. (45)	MRI, T1CE+Flair+DWI	MGMT	105 (74 + 31)	3051	13	0.90^{a}	Correlation coefficient, MRMR, Logistic regression
Li et al. (46)	MRI, T1+T1CE+T2+Flair	MGMT	193 (133 + 60)	1705	6	0.88^{a}	Correlation coefficient, RF-Boruta, RF
Qian et al. (47)	PET	MGMT	86 (59 + 27)	1450	3	80%^{b}	Correlation coefficient, RF Extra trees, SVM, Neural network
Su et al. (48)	MRI, T1+T1CE+T2+Flair	H3 K27M	100 (75 + 25)	85	10	0.85^{a}	TPOT
Choi et al. (15)	MRI, T1CE+T2+Flair	IDH1/2 status	1166 (727 + 439)	NA	NA	0.96^{a}	CNNs, RF
Jin et al. (38)	H&E Pathological slides	IDH1/2 and 1p/19q	323 (267 + 56)	NA	NA	87.6%^{b}	CNNs
Liu et al. (49)	H&E Pathological slides	IDH1/2 status	200 (6:1:1)	NA	NA	0.931^{a}	CNNs

CNNs, Convolutional neural networks; DWI, Diffusion Weighted Imaging; Flair, Fluid-attenuated inversion recovery; H&E, haematoxylin and eosin; IDH, Isocitrate dehydrogenase; Lasso, Least absolute shrinkage and selection operator; MGMT, O (6)-methylguanine-DNA methyltransferase; MRI, Magnetic resonance imaging; MRMR, Minimum redundancy and maximum relevance algorithm; NA, Not available; PET, Positron emission tomography; 1p19q, the co-deletion status of the 1p/19q chromosome arms; RF, Random forest; SVM, Support vector machines; T1, T1-weighted MRI; T2, T2-weighted MRI; T1CE, contrast-enhanced T1-weighted MRI; TPOT, the Tree-based Pipeline Optimization Tool.

The values in the performance column were achieved using the best model in the test set. ^{a} and ^{b} are used to mark the AUC and accuracy values, respectively.

be divided into several directions. On the one hand, researchers can first identify the prognostic imaging features (like OS or progression-free survival, PFS) to calculate individualized indicators (such as radiomic risk score, RRS), based on which patients will be further classified into different subgroups (low- or high-risk groups). Some bioinformatics analysis methods such as Gene Ontology (GO), Gene Set Enrichment Analysis (GSEA) analysis can be further used to detect differences in molecular biology of different subgroups, corresponding signal pathways, and downstream biological processes (28, 54, 55). Another research paradigm is the reverse process of the above method (13, 56). Bioinformatics analysis can reveal the pathways, biological processes, and expression of protein corresponding to specific molecular markers detected by sequencing technique. These biological processes can finally be mapped in the form of different imaging phenotypes (like necrosis or edema) on multi-modal images, and then captured by the quantitative radiomic features (such as shape, texture, or wavelet). A pioneering radiogenomics study by Zinn et al. (13) showed that radiomic-features could significantly predict periostin expression status in GB patients (Area under the receiver operating characteristic curve, AUC=76.56%) and orthotopic xenografts (AUC=92.26%). The authors tried to establish causality between radiomic texture features and expression levels in a pre-clinical model. This was one of the first radiogenomics studies highlighting a causal linkage between specific gene expression status and radiomic-features of GB. This kind of research paradigm is more suitable for the verification of the radiogenomics model and its corresponding conclusions. However, to date, it has not been elucidated how specific biological processes reflect phenotypes through imaging. Additional multidisciplinary studies in the future are needed to shed more light on this subject. **Figure 2** describes the common steps of an imaging-genomic study and **Table 2** summarizes several imaging-genomic studies in patients with GB.

Although imaging-genomics has broad application prospects, current researches mainly focus on one or several aspects of the procedures and purpose discussed above. Here, we will review the recent progress in the utilization of the imaging-genomics analysis in GB patients.

PROGRESS OF IMAGING-GENOMICS IN GB

Candidate Molecular Alterations

Isocitrate Dehydrogenase 1 and 1p/19q Codeletion

Genetic analysis revealed that the IDH1 gene mutations exist in approximately 12% of GB (21). Tumor-related IDH1 mutations result in the enhancement of enzyme function, which causes the accumulation of R(-)-2-hydroxyglutarate (2HG), and the accumulation of 2HG might drive oncogenesis (60). IDH1 mutation is also an independent factor for predicting the prognosis of patients with glioma (61), and the IDH and 1p/19q codeletion status has been introduced into the revised WHO classification of CNS tumors (6).

There have been numerous studies attempted to reveal the links between the IDH1 and 1p/19q codeletion status with the imaging-phenotypic appearance of GB (15, 38, 41–43). A model for predicting the state of molecular markers (such as IDH1) needs to take into account the distribution of data, considering that the vast majority of GB are IDH1 wild types (17). In this context, some techniques (like the synthetic minority oversampling technique (SMOTE) and adaptive synthetic) have been proven to have better performance in some cases (62). In a study of 225 GB patients by Li et al. (41), after balancing the data with the SMOTE algorithm, the model combining all-region (necrosis, enhancement, non-enhancement, and edema areas) radiomics features with clinical information achieved the best performance of a 97% accuracy. The radiomics model with multi-regional features has the potential to predict the IDH1 mutation status of GB patients, and the data balance is beneficial to improve the classification performance. Li and colleagues (44) used 18F-fluorodeoxyglucos PET scans from 127 patients, after feature selection through least absolute shrinkage and selection operator (LASSO), the final SVM model used 11 features and yielded an area under curve (AUC) of 90% in the validation set. In general, radiomics utilizing multimodal images seems to achieve better performance than a single modality. Nevertheless, there are also exceptions. Using a 3D-Dense-UNets approach, Yogananda et al. (63) reported a mean accuracy of 97.14% in predicting IDH status. Interestingly, the authors demonstrated similar performance when comparing the T2-based network with the multi-contrast network. One possible reason is that the deep-learning networks using conventional single-modality MR images avoid the effect of head motion due to long scan times, which helps promote clinical translation.

Choi et al. combined perfusion and conventional MRI for the prediction of IDH genotype in a group of 463 patients. The authors reported that the long short-term memory model, a type of recurrent neural network (RNN), showed performance with an accuracy of 92.8% in the validation set and 91.7% in the test set (50). Also, the RNN model can provide certain interpretability by demonstrating which temporal features are crucial for the prediction of IDH genotypes. By performing principal component analysis on their final CNN layer, Chang and co-workers (42) selected the imaging features and obtained a prediction accuracy of 94% for IDH status and 92% for 1p19q codeletion. Recently, Choi and colleagues (15) used MRI data from multiple centers to predict IDH genotypes in a cohort of 1,166 patients with glioma (WHO grade II-IV). Researchers developed an automated approach containing two CNN models, one for tumor segmentation and the other for IDH status prediction. Finally, the automated model achieved the highest diagnostic accuracy of 93.8% and 87.9% in the internal and external test set, respectively. A fully automated process is conducive to the repeatability and generalization of results, which reduces the interference of human factors and is beneficial to future application and promotion of the models.

O (6)-Methylguanine-DNA Methyltransferase

The MGMT gene encodes a DNA repair enzyme named O6-alkylguanine DNA alkyltransferase (64). In this process, the methylmoiety is transferred onto the MGMT protein and thus

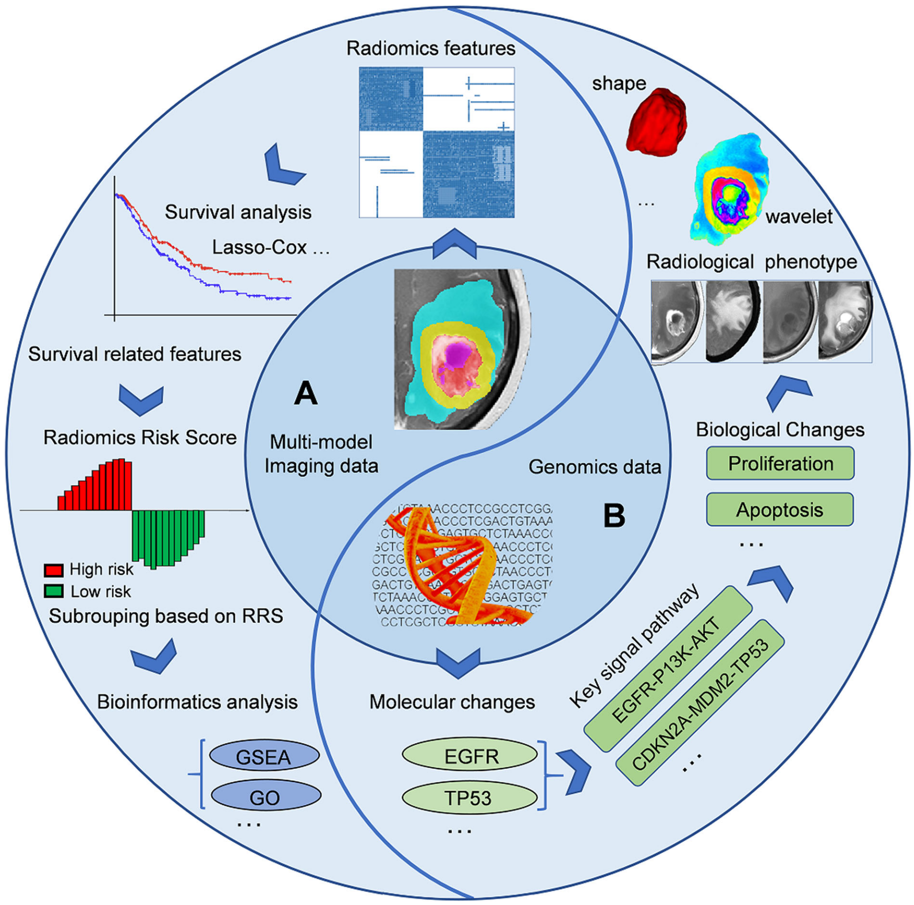

Radiomics features
shape
...
wavelet
Radiological phenotype
Survival analysis
Lasso-Cox ...
Survival related features
Radiomics Risk Score
A
Multi-model
Imaging data
Genomics data
B
Biological Changes
Proliferation
Apoptosis
...
High risk
Low risk
Subrouping based on RRS
Key signal pathway
EGFR-P13K-AKT
CDKN2A-MDM2-TP53
...
Bioinformatics analysis
Molecular changes
GSEA
GO
...
EGFR
TP53
...

FIGURE 2 | General workflow of imaging-genomics studies. A imaging-genomics research can generally build the relationship between imaging phenotypes and genetic characteristics from two perspectives. **(A)**, Left semicircle, from imaging phenotypes to genomic characteristics: using features extracted from the sub-areas of tumors in multi-modal images to construct individualized RRS or divide patients into different subgroups, and utilize the corresponding sequencing data for further bioinformatics analysis (such as GO or GSEA). **(B)**, Right semicircle, from genomic characteristics to imaging phenotypes: bioinformatics analysis can reveal the pathways, biological processes, and expression of protein corresponding to specific molecular markers detected by sequencing technique. These biological processes can finally be mapped in the form of different imaging phenotypes (like necrosis or edema) on multi-modal images, and then captured by the quantitative radiomic features (such as shape, texture, or wavelet features). EGFR, Epidermal growth factor receptor; GO, Gene Ontology; GSEA, Gene set enrichment analysis; RRS, Radiomic risk score.

is consumed (64, 65). The epigenetic modification of the cytosine-phosphate-guanine (CpG) island at specific CpG sites of the MGMT promoter silences the gene, resulting in low DNA alkylation repair efficiency and enhanced response to temozolomide (TMZ) (66, 67).

The methylation status of the MGMT gene promoter has been considered as a crucial biomarker of tumor response to TMZ chemotherapy. Consequently, it has become an important molecular marker in imaging-genomics research (68–71). Sasaki and colleagues (72) extracted features from MRI data of 201 newly diagnosed GB patients. The diagnostic model of MGMT promoter using two significant Gray-level Co-occurrence Matrix texture features selected by LASSO only yielded a low accuracy of 67%, which remains insufficient for practical use. Wei and colleagues (45) found that the fusion radiomics signature (clinical factors and texture features) showed great performance for predicting MGMT promoter status of glioma, with an AUC value of 0.902 in the validation cohort. Moreover, the model could classify patients into low-risk and high-risk groups for overall survival (OS) after TMZ chemotherapy, which may serve as a basis for individualized treatment recommendations. Li et al. (46) predicted the MGMT promoter methylation used 1,705 radiomics features from multiregional and multiparametric MRI in multicenter cohorts. The final model achieved an accuracy of 80% in the test dataset. Distinguished from the studies in the past (43), however, this study suggested that combing clinical factors with radiomics features did not improve the prediction performance.

PET-based radiomics is also a promising method to noninvasively evaluate the MGMT status of gliomas. Qian et al. (47) extracted features from PET images of 86 patients with GB. Features were selected by comparing the weighted feature coefficient and filtering with bivariate analysis. The Random Forest model using the appropriate parameters from

TABLE 2 | Applications of imaging-genomics for exploring potential signaling pathway or biological processes or in GB.

Study	Imaging Modality	General Purpose	Total Patients Training + Testing	No. of features Initial + Final		Performance AUC/ Accuracy	Statistical Analysis
Choi et al. (51)	MRI, T1+T1CE+T2 +Flair	Exploring biological characteristics behind imaging phenotypes in GB.	200 (144 + 56)	478	7	NA	Lasso-Cox, RRS, GSEA
Itakura et al (57).	MRI, T1CE	Verifying distinct pathway changes behind different imaging subtypes of GB.	265 (121 + 144)	388	388	NA	k-means consensus clustering, IGP
Beig et al. (28)	MRI, T1CE+T2 +Flair	Exploring prognostic stratification-based biological processes in GB.	203 (130 + 73)	2850	25	0.84^c	Lasso-Cox, RRS, GO, GSEA
Park et al. (29)	MRI,T1+T1CE +T2DWI +Flair +DSC	Identifying Core signaling pathway in GB. (RTK, p53, and RB1 pathways).	120 (85 + 35)	6472	123	(RTK) 0.88^a	Lasso, RF, logistic regression,
Akbari et al. (58)	MRI, T1+T1CE+T2 +Flair	Predicting status of EGFRvIII and exploring associated biological processes.	129 (75 + 54)	NA	NA	0.86^a	Multivariate model, SVM
Zinn et al. (52)	MRI, T1+T1CE +Flair	TP53-PTEN-EGFR mutational landscape in GB.	29	2480	457	NA	Spearman rank correlation Hierarchical clustering
Beig et al. (53)	MRI, T1CE+ T2 +Flair	Predicting hypoxia pathway and overall survival in GB.	115 (85 + 30)	270	8	0.83^c	GSEA, Cox model, Unsupervised clustering
Hsu et al. (59)	MRI, T1CE+ DWI	Classifying immunophenotypes of GB by radiomic features.	116 (32 + 84)	9809	76	79%^{b}	RF, Information gain, GSEA Logistic regression
Beig et al. (54)	MRI, T1CE+ T2 +Flair	Identifying sex-specific biological processes and prognostic of GB.	313 (130 + 183)	2850	8	0.88^c	Lasso-Cox, RRS, GO, GSEA
Zinn et al. (13)	MRI, T1CE+ T2 +Flair	Validating causality between Periostin expression status and MRI-features in GB.	93 (xenografts=40)	4880	31	0.93^a (xenografts)	Lasso, GSEA, XGboost

DSC, Dynamic susceptibility contrast; DWI, Diffusion weighted imaging; EGFR, Epidermal growth factor receptor; Flair, Fluid-attenuated inversion recovery; GO, Gene ontology; GSEA, Gene set enrichment analysis; IGP, in-group proportion statistic; Lasso, Least absolute shrinkage and selection operator; MRI, Magnetic resonance imaging; NA, Not available; PET, Positron emission tomography; RB1, Retinoblastoma 1; RF, Random forest; RRS, Radiomic risk score; RTK, Receptor tyrosine kinase; SVM, Support vector machines; T1, T1-weighted MRI; T2, T2-weighted MRI; T1CE, contrast-enhanced T1-weighted MRI; XGboost, Extreme Gradient Boosting.

The values in the performance column were achieved using the best model in the test set. ^{a}, ^{b}, and ^{c} are used to mark the AUC, accuracy, and concordance index values, respectively.

the grid search achieved 80% ± 10% accuracy for the prediction of MGMT promoter status. Kong et al. (73) used PET images from 107 patients with WHO grade II-IV primary diffuse gliomas for prediction. Using Wilcoxon ranksum test for features selection, a radiomics signature comprising five final features for the prediction of MGMT promoter status could be identified, yielding an AUC of 0.86 in the validation sets. However, this model lacks further validation on an external dataset.

The CNN-based classifiers achieved the accuracy from 82.7% to 94.9% in predicting the methylation status of MGMT promoter in GB (74, 75). Chen et al. (74) utilized a deep learning pipeline for automatic tumor segmentation and MGMT promoter status prediction in an end-to-end manner for GB patients. The better tumor segmentation and MGMT prediction performance both came from Fluid-attenuated inversion recovery (FLAIR) images. The CNN-based prediction model in this study had only 4 layers and achieved an accuracy of 82.7%, which might be further improved by a deeper network architecture with multi-layers. Korfiatis et al. (75) compared three different residual deep neural network (ResNet) architectures to evaluate their performance in predicting MGMT promoter status in patients with GB. The researchers found that the ResNet50 architecture (the deepest model with 50 layers) outperformed the two shallower architectures (ResNet18 with 18 layers, ResNet34 with 34 layers), and achieved an accuracy of 94.9% in the test set. Networks with deeper architectures could potentially improve the performance of the model, which also increases the risk of overfitting. Transfer learning and fine-tuning of the existing pre-trained network are alternative methods to improve the utilization of the limited data (34).

Other Potential Genetic/Molecular Markers

EGFR is a transmembrane tyrosine kinase that regulates normal cellular growth in epithelial cell lines. The most common EGFR mutant in GB is EGFRvIII, which is found in 31% of patients (76). Using support vector machine–based approaches, Akbari et al. (58) construct an imaging signature of EGFRvIII and achieved an accuracy of 87% in 129 patients with primary GB. The authors also reported that EGFRvIII mutant GB exhibit increased neovascularization and cell density, which is consistent with a more infiltrative phenotype. Zinn and colleagues (52) demonstrated a relationship between MRI radiomics signatures and the TP53-PTEN-EGFR mutational landscape in 29 GB patients from The Cancer Genome Atlas (TCGA). The radiomics features generated by texture analysis revealed different cellular biofunctions, such as angiogenesis, invasion, and immune response. Despite some gaps exist, glioma radiomics landscape is approaching its genomic counterpart in heterogeneity and complexity.

Diffuse midline glioma with histone H3-K27M mutation is a newly defined entity in the WHO group of grade IV diffuse gliomas (77). Those tumors carry a poor prognosis and frequently occur in brainstem, thalamus in the pediatric population. The underlying histological and radiographic variations (78) make it possible to predict the H3-K27M status of a tumor using radiomics analysis. Su and colleagues (48) performed an automated machine learning analysis on MRI data from 100 patients with midline gliomas. After feature selection and model optimization using The Tree-based Pipeline Optimization Tool (TPOT), they finally obtained an AUC of 0.903 and an average precision of 0.911 for the prediction of H3 K27M mutant status in midline glioma. Automated machine learning methods (such as TPOT) are relatively researcher-independent, which increases the robustness of the results.

Signaling Pathways and Biological Processes

Genomic analysis has identified a highly interconnected network of aberrations in GB, including three major pathways: receptor tyrosine kinase (RTK), the retinoblastoma 1 (RB), and tumor protein p53 tumor suppressor pathways (4). Radiogenomics studies have revealed the relationship between different imaging features and certain signaling pathways or biological processes.

In a study of 120 patients with GB, using diffusion, perfusion MRI, and next-generation sequencing technology, Park and colleague (29) combined radiomic and genomic features to predict core signaling pathways in IDH wild-type GB. Different types of radiomic features were identified by the combined radiogenomics model, corresponding to RTK, P53, and RB pathways respectively. This study demonstrates that MRI features can help noninvasively identify the core signaling pathway alterations in GB, which may further support the use of targeted therapy for glioblastoma rather than a one-size-fits-all approach. Another study based on perfusion MRI showed that the elevated perfusion features were significantly associated with poor patient survival in a subgroup of GB patients. Angiogenesis and hypoxia pathways were enriched in this subgroup, suggesting the possible efficacy of anti-angiogenic therapy (79). Similarly, imaging features from enhancing tumor and edematous regions were identified to be associated with the hypoxia pathway (53), imaging markers of hypoxia can discriminate GB patients as short, mid, and long-term survivors.

Radiogenomics studies also highlight sex-related biological variations in GB patients. Colen et al. (80) revealed sex-specific molecular mechanisms for cell death (MYC oncogenic in female GB patients, TP53 apoptotic in male GB patients). Similarly, a recent study by Beig et al. (54) suggested that higher expression of texture features from enhancing tumor regions seemed to be more enriched in 'high-risk' group in the male population. This result appeared to be opposite in the female population, in which the same features were enriched in the 'low-risk' group. These findings suggesting the importance of treating sex as a covariate in the design of a radiogenomics study.

Heterogeneity and Molecular Subtype

Intra-tumoral Heterogeneity in GB has been a topic of long-standing interest (23, 81). Spatio-temporally comprehensive assessment of the tumor heterogeneity requires multiple spatially sampling at different time points. In daily clinical practice, effective tissue sampling remains a significant challenge in accurately evaluating the intratumoral heterogeneity of a GB. Radiogenomics has shown the initial prospects for spatially evaluating the regional and genetically distinct subtypes that coexist within a single GB tumor (30, 82). By collected 48 image-guided biopsies from 13 GBs, Hu et al. (30) co-registered each biopsy location with MRI texture map to correlate spatially matched imaging measurements with regional genetic status. They identified significant imaging correlations for several driver genes, including EGFR, PTEN, RB1, TP53, and more. The author also emphasized that even within one tumor segment (such as the enhancement part in T1 contrast-enhanced MRI), regional genetic diversity can also exist. This suggests the need to improve the image-based assessment of genetic heterogeneity beyond the use of the subregion presented in a single imaging mode.

Images contain information about tumor phenotypes that are regulated not only by cell-intrinsic biological processes but also by the tumor microenvironment (59, 83–85). Recently, machine learning-based radiomics models demonstrated the capacity to evaluate enrichment levels of different immune cells in GB patients. Using radiomics features of MR images, the radiomics models could classify the immunophenotypes of GB and can predict patient prognosis (59). Lin et al. (84) used a consistent clustering method to classify radiological features into two subgroups. The two subtypes have significantly different histological stages and molecular factors (such as IDH and MGMT). Furthermore, biological information analysis hinted that the inferior prognosis subtype may be more responsive to immunotherapy. At the same time, if GB patients receive immunotherapy (such as immune checkpoint blockade therapy), the responses of therapy might also be distinguished by radiomics signatures noninvasively (86). Cho et al. (83) revealed that immune cell markers have significant correlations with perfusion and diffusion MRI features in a study of 60 GB patients. The apparent diffusion coefficient values were correlated with the expression level of immune cell markers, which can be used as an immune biomarker to predict the progression and prognosis of GB patients. It should be noted that many classification methods based on radiomics characteristics are still in an early developmental stage. Although these potential associations are encouraging, some conclusions still need further verification.

Prognostic Stratification-Based Tumor Biological Characteristics

A large number of retrospective studies have revealed the link between the radiomics characteristics and survival time of GB patients (70, 87). According to radiomics risk stratification, some studies (28, 54, 84, 88) also tried to explore the corresponding tumor biological information. Based on survival-related imaging features, some studies (28, 54) construct individualized RRS, by which the patients will be divided into subgroups with different risk stratifications. Further, bioinformatics analysis such as GO and/or GSEA will be used to identify the signaling pathway networks or specific biological characteristics that were associated with RRS features. For example, using data from public platforms, Beig and colleagues (28) extracted 936

3D-radiomics features from T1, T2, and Flair MR images of 203 patients with GB. Finally, the GO and GSEA revealed associations of RRS signatures with signaling pathways for cell adhesion, cell differentiation, and angiogenesis. A somewhat similar approach was followed by Choi et al. (51) who used RRS weighted by Lasso-Cox model to classify patients with GB. Finally, GSEA suggested that the transcriptomic characteristics enriched in personalized subtypes were consistent with radiomics phenotypes. For example, the rim-enhancing necrotic subtype could be described by T2-derived radiomic features and further highlighted by the inflammatory genomic signatures.

Radiogenomics has the potential to unravel the molecular underpinning behind the imaging-derived phenotypes. However, these associations seem to be indirect, since they were constructed through survival variables, such as OS and PFS. Indirect associations may obscure or expand the relationships between radiomics characteristics and molecular phenotypes. In addition to using optimized AI models to integrate all such information, radiogenomics biological validation in animal models (13) would be a promising direction in the future.

Personalized Treatment and Assessment

At present, several studies are exploring combination imaging-genomics with the assessment of early treatment response. Imaging biomarkers have shown great potential to stratify the therapeutic response of patients with recurrent glioblastoma treated with bevacizumab (89). EGFR extracellular domain mutations in GB also present opportunities for imaging and therapeutic development (56). Wei and colleagues (45) used logistic regression to generate a fusion radiomics signature based on conventional MRI of 105 patients with WHO grade II-IV astrocytomas. The fusion signature exhibited great power for predicting the methylation status of MGMT promoter and evaluating temozolomide (TMZ) chemotherapy effect. Petrova et al. (90) collected perfusion and diffusion images from 54 subjects with recurrent GB and subsequently treated with bevacizumab. Among the six different classifiers, the authors found that SVM was able to identify 97% of subjects that would respond to bevacizumab with accuracy for PFS and OS of 82% and 78%. This can help assess whether a GB patient can benefit from anti-angiogenic therapy. However, the model still needs to prove its repeatability and generalizability in an extensive cohort study.

GB is a heterogeneous disease composed of multiple molecular subtypes with different clinical courses, leading to different treatment strategies and prognoses. Radiomics and genomics attempt to explain the diversity of tumors from different perspectives. A more fused model might be used to integrate clinical information and all experimental data of an individual, and propose corresponding molecular networks specific to a personalized GB for efficient early-stage diagnosis, precise treatment recommendations, and prognostic assessment. At present, the underlying transformation of imaging-genomics remains only at an early stage of the investigation. Imaging signatures identified by animal models may be useful for future clinical trials enriched for patients with specific subtypes of GB (13, 86).

FUTURE CHALLENGES

Big Data and Data Sharing

One of the fundamental challenges in imaging-genomics research is the availability of large, standardized, well-annotated data sets. Many studies lacked external testing for evaluating the performance of the models due to the limit of data. While external testing is crucial to confirm model generalizability across different institutions. Data sharing between institutes would bring new opportunities for researchers. The Cancer Imaging Archive (TCIA, https://www.cancerimagingarchive.net/) (91), a part of the Cancer Genome Atlas (TCGA), can provide open-access cancer medical imaging and corresponding genome resources to support imaging-genomics researches (59, 63). Some information about the tumors and corresponding supporting data has been summarized on the TCIA website (https://www.cancerimagingarchive.net/collections/).

Radiological images and biological data from CPTAC (Clinical Proteomic Tumor Analysis Consortium, https://proteomics.cancer.gov/programs/cptac) (92, 93) have also been made publicly available to enable researchers to investigate the imaging phenotypes which may correlate to corresponding genomic, proteomic, and clinical data (28, 54). Additionally, the Ivy Glioblastoma Atlas Project (IvyGAP, http://glioblastoma.alleninstitute.org/) (94) provides resources for exploring the anatomic and genetic basis of GB at the radiological and molecular levels. The dataset contains RNA sequencing data, tissue sections, as well as partial MR data in TCIA. Data from these public databases has motivated and continues to motivate the technological exploration and development in imaging-genomics field. However, such data are far from sufficient. Further researches focused on developing improved algorithms (such as transfer learning or data augmentation technologies) optimized for smaller datasets are also needed.

Quality Control, Interpretability, and Application Challenge

Even for the same data, two different calculation software might generate different feature values (95). The lack of standardized definitions will hamper the clinical use of the models. Undoubtedly, quality control and repeatability should run through the entire imaging genomic study, including data collection, preprocessing, delineation of ROIs, features extraction and selection, model building, and verification. Lambin et al. (10) have proposed the radiomics quality score (RQS) to assess the quality of radiomics studies, which plays a guiding role in improving the initial design of an imaging-genomics research. The Image Biomarker Standardization Initiative (IBSI) also produced and validated a series of consensus-based reference values for radiomics features (95). These reference values can be used to check whether a radiomics software complies with IBSI standards. Improved research quality and reproducibility can promote the clinical translation of radiomics.

In the clinical setting, there are still many obstacles to the standardized processing of data and the effective deployment of the pre-trained models. The lack of sufficient verification of a model will inevitably limit its further application. Additionally, it is also important to note that features and models generated by AI (especially deep learning) often lack clear interpretable parameters. Even if AI algorithms provide good predictive results, the models and deep features are difficult for clinicians to interpret, which limits the integration of radiomics into routine clinical practice. In the context of a regulated healthcare environment, this "black box" problem of AI algorithms is particularly outstanding. Visualization of models and features is an alternative way to alleviate this problem. Interpretability in AI is an emerging field of increased research efforts in the data science community.

Multi-Omics Integration

Multi-omics is undergoing the big data revolution. Genomic analysis such as gene mutation or altered gene expression alone does not always reflect the real change of the corresponding protein and downstream metabolite. The intricate connections between tumor biology and holistic phenotype could have been better captured by multidimensional data and multi-omics models. In recent years, leveraging AI and digital pathology has shown significant promise in cancer diagnosis (38, 49). Remarkably, there have been attempts to integrate pathological section data and radiological data (96), which is called a new term as radopathomics. The fusion signature combining information from distinct dimensions could better predict discrepancies of treatment response. These results have exhibited the great potential of radiomics in the integration of modern precision medicine.

Predictably, the evolution of the radiomics field will gradually differentiate into "radio-…omics", such as radio-proteomics, radiopatho-genomic, or even radio-metabonomics in the coming future, dependent on the source used. As such, current imaging-genomic association maps entail integrations of large, "multi-omic" data sets with imaging to bridge the missing links. Certainly, such a large number of omics features can easily lead to overfitting. Optimal feature reduction and selection methods and appropriate model validation are also needed to deal with the problem of "Curse of Dimensionality".

CONCLUSION

The evolution of the field of radiomics and radiogenomics is similar to the emergence of the microscope in pathology. It is foreseeable that based on multi-parameter imaging features, more biological information can be mined to evaluate the biological characteristics of GB for supporting individualized patients management. However, many challenges still exist and much work needs to be done. Better interpretability and validation of these tools will make imaging genomics more acceptable in this field. Even though the translation of these models from research to clinical practice is still in progress, they provide promise for guiding precision medicine and subsequent individualized therapeutic intervention.

AUTHOR CONTRIBUTIONS

HL and WZ conceived this review, organized, and critically revised the manuscript. DL and JC did major work of the manuscript. XH, KY, YL, GH, and HG made contributions to the writing and revising of the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This study was supported by grants from the National Natural Science Foundation of China (No. 81972350, 81701671), the project of Jiangsu Provincial Medical Youth Talent (No. QNRC2016047), the Medical Scientific and Technologic Development Project of Nanjing (No. ZKX15035), the Jiangsu Provincial Medical Innovation Team (No. CXTDA2017050), the Medical Science and Technology Development Foundation of Nanjing (No. ZDX16011), and the Nanjing Medical Science and technique Development Foundation (No. QRX17087).

REFERENCES

1. Ostrom QT, Cioffi G, Gittleman H, Patil N, Waite K, Kruchko C, et al. Cbtrus Statistical Report: Primary Brain and Other Central Nervous System Tumors Diagnosed in the United States in 2012-2016. *Neuro-oncology* (2019) 21:v1–v100. doi: 10.1093/neuonc/noz150
2. Glas M, Rath BH, Simon M, Reinartz R, Schramme A, Trageser D, et al. Residual Tumor Cells are Unique Cellular Targets in Glioblastoma. *Ann Neurol* (2010) 68(2):264–9. doi: 10.1002/ana.22036
3. Stupp R, Taillibert S, Kanner A, Read W, Steinberg D, Lhermitte B, et al. Effect of Tumor-Treating Fields Plus Maintenance Temozolomide vs Maintenance Temozolomide Alone on Survival in Patients With Glioblastoma: A Randomized Clinical Trial. *JAMA* (2017) 318(23):2306–16. doi: 10.1001/jama.2017.18718
4. Chin L, Meyerson M, Aldape K, Bigner D, Mikkelsen T, Vandenberg S, et al. Comprehensive Genomic Characterization Defines Human Glioblastoma Genes and Core Pathways. *Nature* (2008) 455(7216):1061–8. doi: 10.1038/nature07385
5. Pearson JRD, Regad T. Targeting Cellular Pathways in Glioblastoma Multiforme. *Signal Transduct Target Ther* (2017) 2:17040. doi: 10.1038/sigtrans.2017.40
6. Louis DN, Perry A, Reifenberger G, Von Deimling A, Figarella-Branger D, Cavenee WK, et al. The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a Summary. *Acta Neuropathol* (2016) 131(6):803–20. doi: 10.1007/s00401-016-1545-1
7. Wang Q, Hu B, Hu X, Kim H, Squatrito M, Scarpace L, et al. Tumor Evolution of Glioma-Intrinsic Gene Expression Subtypes Associates With Immunological Changes in the Microenvironment. *Cancer Cell* (2017) 32(1):42–56.e6. doi: 10.1016/j.ccell.2017.06.003

8. Villanueva-Meyer JE, Mabray MC, Cha S. Current Clinical Brain Tumor Imaging. *Neurosurgery* (2017) 81(3):397–415. doi: 10.1093/neuros/nyx103
9. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, Van Stiphout Rg, Granton P, et al. Radiomics: Extracting More Information From Medical Images Using Advanced Feature Analysis. *Eur J Cancer (Oxford Engl 1990)* (2012) 48(4):441–6. doi: 10.1016/j.ejca.2011.11.036
10. Lambin P, Leijenaar R, Deist TM, Peerlings J, De Jong E, Van Timmeren J, et al. Radiomics: The Bridge Between Medical Imaging and Personalized Medicine. *Nat Rev Clin Oncol* (2017) 14(12):749–62. doi: 10.1038/nrclinonc.2017.141
11. Kuo MD, Jamshidi N. Behind the Numbers: Decoding Molecular Phenotypes With Radiogenomics–Guiding Principles and Technical Considerations. *Radiology* (2014) 270(2):320–5. doi: 10.1148/radiol.13132195
12. Ellingson BM. Radiogenomics and Imaging Phenotypes in Glioblastoma: Novel Observations and Correlation With Molecular Characteristics. *Curr Neurol Neurosci Rep* (2015) 15(1):506. doi: 10.1007/s11910-014-0506-0
13. Zinn PO, Singh SK, Kotrotsou A, Hassan I, Thomas G, Luedi MM, et al. A Coclinical Radiogenomic Validation Study: Conserved Magnetic Resonance Radiomic Appearance of Periostin-Expressing Glioblastoma in Patients and Xenograft Models. *Clin Cancer Res* (2018) 24(24):6288–99. doi: 10.1158/1078-0432.CCR-17-3420
14. Mazurowski MA. Radiogenomics: What it is and Why it is Important. *J Am Coll Radiol JACR* (2015) 12(8):862–6. doi: 10.1016/j.jacr.2015.04.019
15. Choi YS, Bae S, Chang JH, Kang SG, Kim SH, Kim J, et al. Fully Automated Hybrid Approach to Predict the IDH Mutation Status of Gliomas Via Deep Learning and Radiomics. *Neuro Oncol* (2020) 23(2):304–13. doi: 10.1093/neuonc/noaa177 noaa177.
16. Fukuma R, Yanagisawa T, Kinoshita M, Shinozaki T, Arita H, Kawaguchi A, et al. Prediction of IDH and TERT Promoter Mutations in Low-Grade Glioma From Magnetic Resonance Images Using a Convolutional Neural Network. *Sci Rep* (2019) 9(1):20311. doi: 10.1038/s41598-019-56767-3
17. Ohgaki H, Kleihues P. The Definition of Primary and Secondary Glioblastoma. *Clin Cancer Res* (2013) 19(4):764–72. doi: 10.1158/1078-0432.CCR-12-3002
18. Lai A, Kharbanda S, Pope WB, Tran A, Solis OE, Peale F, et al. Evidence for Sequenced Molecular Evolution of IDH1 Mutant Glioblastoma From a Distinct Cell of Origin. *J Clin Oncol* (2011) 29(34):4482–90. doi: 10.1200/jco.2010.33.8715
19. Aldape K, Zadeh G, Mansouri S, Reifenberger G, Von Deimling A. Glioblastoma: Pathology, Molecular Mechanisms and Markers. *Acta Neuropathol* (2015) 129(6):829–48. doi: 10.1007/s00401-015-1432-1
20. Brennan CW, Verhaak RG, Mckenna A, Campos B, Noushmehr H, Salama SR, et al. The Somatic Genomic Landscape of Glioblastoma. *Cell* (2013) 155 (2):462–77. doi: 10.1016/j.cell.2013.09.034
21. Parsons DW, Jones S, Zhang X, Lin JC, Leary RJ, Angenendt P, et al. An Integrated Genomic Analysis of Human Glioblastoma Multiforme. *Science* (2008) 321(5897):1807–12. doi: 10.1126/science.1164382
22. Turner NC, Reis-Filho JS. Genetic Heterogeneity and Cancer Drug Resistance. *Lancet Oncol* (2012) 13(4):e178–85. doi: 10.1016/s1470-2045(11)70335-7
23. Sottoriva A, Spiteri I, Piccirillo SG, Touloumis A, Collins VP, Marioni JC, et al. Intratumor Heterogeneity in Human Glioblastoma Reflects Cancer Evolutionary Dynamics. *Proc Natl Acad Sci USA* (2013) 110(10):4009–14. doi: 10.1073/pnas.1219747110
24. Wang J, Cazzato E, Ladewig E, Frattini V, Rosenbloom DI, Zairis S, et al. Clonal Evolution of Glioblastoma Under Therapy. *Nat Genet* (2016) 48 (7):768–76. doi: 10.1038/ng.3590
25. Francis JM, Zhang CZ, Maire CL, Jung J, Manzo VE, Adalsteinsson VA, et al. EGFR Variant Heterogeneity in Glioblastoma Resolved Through Single-Nucleus Sequencing. *Cancer Discovery* (2014) 4(8):956–71. doi: 10.1158/2159-8290.CD-13-0879
26. Köber V, Yang J, Barah P, Wu Y, Stichel D, Gu Z, et al. Evolutionary Trajectories of IDH(WT) Glioblastomas Reveal a Common Path of Early Tumorigenesis Instigated Years Ahead of Initial Diagnosis. *Cancer Cell* (2019) 35(4):692–704.e12. doi: 10.1016/j.ccell.2019.02.007
27. Agarwal S, Sane R, Oberoi R, Ohlfest JR, Elmquist WF. Delivery of Molecularly Targeted Therapy to Malignant Glioma, a Disease of the Whole Brain. *Expert Rev Mol Med* (2011) 13:e17. doi: 10.1017/s1462399411001888
28. Beig N, Bera K, Prasanna P, Antunes J, Correa R, Singh S, et al. Radiogenomic-Based Survival Risk Stratification of Tumor Habitat on Gd-T1w MRI is Associated With Biological Processes in Glioblastoma. *Clin Cancer Res* (2020) 26(8):1866–76. doi: 10.1158/1078-0432.ccr-19-2556
29. Park JE, Kim HS, Park SY, Nam SJ, Chun SM, Jo Y, et al. Prediction of Core Signaling Pathway by Using Diffusion- and Perfusion-based Mri Radiomics and Next-generation Sequencing in Isocitrate Dehydrogenase Wild-type Glioblastoma. *Radiology* (2020) 294(2):388–97. doi: 10.1148/radiol.2019190913
30. Hu LS, Ning S, Eschbacher JM, Baxter LC, Gaw N, Ranjbar S, et al. Radiogenomics to Characterize Regional Genetic Heterogeneity in Glioblastoma. *Neuro Oncol* (2017) 19(1):128–37. doi: 10.1093/neuonc/now135
31. Lasocki A, Rosenthal MA, Roberts-Thomson SJ, Neal A, Drummond KJ. Neuro-Oncology and Radiogenomics: Time to Integrate? *AJNR Am J Neuroradiol* (2020) 41(11):1982–8. doi: 10.3174/ajnr.A6769
32. Rudie JD, Rauschecker AM, Bryan RN, Davatzikos C, Mohan S. Emerging Applications of Artificial Intelligence in Neuro-Oncology. *Radiology* (2019) 290(3):607–18. doi: 10.1148/radiol.2018181928
33. Lecun Y, Bengio Y, Hinton G. Deep Learning. *Nature* (2015) 521(7553):436–44. doi: 10.1038/nature14539
34. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A Survey on Deep Learning in Medical Image Analysis. *Med Image Anal* (2017) 42:60–88. doi: 10.1016/j.media.2017.07.005
35. Lao J, Chen Y, Li ZC, Li Q, Zhang J, Liu J, et al. A Deep Learning-Based Radiomics Model for Prediction of Survival in Glioblastoma Multiforme. *Sci Rep* (2017) 7(1):10353. doi: 10.1038/s41598-017-10649-8
36. Liu Z, Wang S, Dong D, Wei J, Fang C, Zhou X, et al. The Applications of Radiomics in Precision Diagnosis and Treatment of Oncology: Opportunities and Challenges. *Theranostics* (2019) 9(5):1303–22. doi: 10.7150/thno.30309
37. Kurc T, Bakas S, Ren X, Bagari A, Momeni A, Huang Y, et al. Segmentation and Classification in Digital Pathology for Glioma Research: Challenges and Deep Learning Approaches. *Front Neurosci* (2020) 14:27. doi: 10.3389/fnins.2020.00027
38. Jin L, Shi F, Chun Q, Chen H, Ma Y, Hameed NUF, et al. Artificial Intelligence Neuropathologist for Glioma Classification Using Deep Learning on Hematoxylin and Eosin Stained Slide Images and Molecular Markers. *Neuro Oncol* (2020) 23(1):42–52. doi: 10.1093/neuonc/noaa163 noaa163.
39. Tajbakhsh N, Shin JY, Gurudu SR, Hurst RT, Kendall CB, Gotway MB, et al. Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning? *IEEE Trans Med Imaging* (2016) 35(5):1299–312. doi: 10.1109/tmi.2016.2535302
40. Kermany DS, Goldbaum M, Cai W, Valentim CCS, Liang H, Baxter SL, et al. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. *Cell* (2018) 172(5):1122–1131.e9. doi: 10.1016/j.cell.2018.02.010
41. Li ZC, Bai H, Sun Q, Zhao Y, Lv Y, Zhou J, et al. Multiregional Radiomics Profiling From Multiparametric MRI: Identifying an Imaging Predictor of IDH1 Mutation Status in Glioblastoma. *Cancer Med* (2018) 7(12):5999–6009. doi: 10.1002/cam4.1863
42. Chang P, Grinband J. Deep-Learning Convolutional Neural Networks Accurately Classify Genetic Mutations in Gliomas. *AJNR Am J Neuroradiol* (2018) 39(7):1201–7. doi: 10.3174/ajnr.A5667
43. Zhang B, Chang K, Ramkissoon S, Tanguturi S, Bi WL, Reardon DA, et al. Multimodal MRI Features Predict Isocitrate Dehydrogenase Genotype in High-Grade Gliomas. *Neuro Oncol* (2017) 19(1):109–17. doi: 10.1093/neuonc/now121
44. Li L, Mu W, Wang Y, Liu Z, Liu Z, Wang Y, et al. A Non-invasive Radiomic Method Using (18)F-Fdg PET Predicts Isocitrate Dehydrogenase Genotype and Prognosis in Patients With Glioma. *Front Oncol* (2019) 9:1183. doi: 10.3389/fonc.2019.01183
45. Wei J, Yang G, Hao X, Gu D, Tan Y, Wang X, et al. A Multi-Sequence and Habitat-Based MRI Radiomics Signature for Preoperative Prediction of MGMT Promoter Methylation in Astrocytomas With Prognostic Implication. *Eur Radiol* (2019) 29(2):877–88. doi: 10.1007/s00330-018-5575-z
46. Li ZC, Bai H, Sun Q, Li Q, Liu L, Zou Y, et al. Multiregional Radiomics Features From Multiparametric MRI for Prediction of MGMT Methylation Status in Glioblastoma Multiforme: a Multicentre Study. *Eur Radiol* (2018) 28 (9):3640–50. doi: 10.1007/s00330-017-5302-1

47. Qian J, Herman MG, Brinkmann DH, Laack NN, Kemp BJ, Hunt CH, et al. Prediction of MGMT Status for Glioblastoma Patients Using Radiomics Feature Extraction From 18f-DOPA-PET Imaging. *Int J Radiat OncologyBiologyPhys* (2020) 108(5):1339–46. doi: 10.1016/j.ijrobp.2020.06.073
48. Su X, Chen N, Sun H, Liu Y, Yang X, Wang W, et al. Automated Machine Learning Based on Radiomics Features Predicts H3 K27M Mutation in Midline Gliomas of the Brain. *Neuro Oncol* (2020) 22(3):393–401. doi: 10.1093/neuonc/noz184
49. Liu S, Shah Z, Sav A, Russo C, Berkovsky S, Qian Y, et al. Isocitrate Dehydrogenase (IDH) Status Prediction in Histopathology Images of Gliomas Using Deep Learning. *Sci Rep* (2020) 10(1):7733. doi: 10.1038/s41598-020-64588-y
50. Choi KS, Choi SH, Jeong B. Prediction of IDH Genotype in Gliomas With Dynamic Susceptibility Contrast Perfusion MR Imaging Using an Explainable Recurrent Neural Network. *Neuro Oncol* (2019) 21(9):1197–209. doi: 10.1093/neuonc/noz095
51. Choi SW, Cho HH, Koo H, Cho KR, Nenning KH, Langs G, et al. Multi-Habitat Radiomics Unravels Distinct Phenotypic Subtypes of Glioblastoma With Clinical and Genomic Significance. *Cancers* (2020) 12(7):1707–12. doi: 10.3390/cancers12071707
52. Zinn PO, Singh SK, Kotrotsou A, Abrol S, Thomas G, Mosley J, et al. Distinct Radiomic Phenotypes Define Glioblastoma TP53-PTEN-EGFR Mutational Landscape. *Neurosurgery* (2017) 64(CN_suppl_1):203–10. doi: 10.1093/neuros/nyx316
53. Beig N, Patel J, Prasanna P, Hill V, Gupta A, Correa R, et al. Radiogenomic Analysis of Hypoxia Pathway is Predictive of Overall Survival in Glioblastoma. *Sci Rep* (2018) 8(1):7. doi: 10.1038/s41598-017-18310-0
54. Beig N, Singh S, Bera K, Prasanna P, Singh G, Chen J, et al. Sexually Dimorphic Radiogenomic Models Identify Distinct Imaging and Biological Pathways That are Prognostic of Overall Survival in Glioblastoma. *Neuro Oncol* (2020) 23(2):251–63. doi: 10.1093/neuonc/noaa231 noaa231.
55. Beig N, Bera K, Tiwari P. Introduction to Radiomics and Radiogenomics in Neuro-Oncology: Implications and Challenges. *Neurooncol Adv* (2020) 2 (Suppl 4):iv3–iv14. doi: 10.1093/noajnl/vdaa148
56. Binder ZA, Thorne AH, Bakas S, Wileyto EP, Bilello M, Akbari H, et al. Epidermal Growth Factor Receptor Extracellular Domain Mutations in Glioblastoma Present Opportunities for Clinical Imaging and Therapeutic Development. *Cancer Cell* (2018) 34(1):163–177 e7. doi: 10.1016/j.ccell.2018.06.006
57. Itakura H, Achrol AS, Mitchell LA, Loya JJ, Liu T, Westbroek EM, et al. Magnetic Resonance Image Features Identify Glioblastoma Phenotypic Subtypes With Distinct Molecular Pathway Activities. *Sci Transl Med* (2015) 7(303):303ra138. doi: 10.1126/scitranslmed.aaa7582
58. Akbari H, Bakas S, Pisapia JM, Nasrallah MP, Rozycki M, Martinez-Lage M, et al. In Vivo Evaluation of EGFRvIII Mutation in Primary Glioblastoma Patients Via Complex Multiparametric MRI Signature. *Neuro Oncol* (2018) 20 (8):1068–79. doi: 10.1093/neuonc/noy033
59. Hsu JB, Lee GA, Chang TH, Huang SW, Le NQK, Chen YC, et al. Radiomic Immunophenotyping of GSEA-Assessed Immunophenotypes of Glioblastoma and Its Implications for Prognosis: a Feasibility Study. *Cancers* (2020) 12(10):3039–55. doi: 10.3390/cancers12103039
60. Dang L, White DW, Gross S, Bennett BD, Bittinger MA, Driggers EM, et al. Cancer-Associated IDH1 Mutations Produce 2-Hydroxyglutarate. *Nature* (2010) 465(7300):966. doi: 10.1038/nature09132
61. Sanson M, Marie Y, Paris S, Idbaih A, Laffaire J, Ducray F, et al. Isocitrate Dehydrogenase 1 Codon 132 Mutation is an Important Prognostic Biomarker in Gliomas. *J Clin Oncol* (2009) 27(25):4150–4. doi: 10.1200/jco.2009.21.9832
62. Xie C, Du R, Ho JW, Pang HH, Chiu KW, Lee EY, et al. Effect of Machine Learning Re-Sampling Techniques for Imbalanced Datasets in F-FDG PET-Based Radiomics Model on Prognostication Performance in Cohorts of Head and Neck Cancer Patients. *Eur J Nucl Med Mol Imaging* (2020) 47(12):2826–35. doi: 10.1007/s00259-020-04756-4
63. Bangalore Yogananda CG, Shah BR, Vejdani-Jahromi M, Nalawade SS, Murugesan GK, Yu FF, et al. A Novel Fully Automated MRI-based Deep-Learning Method for Classification of IDH Mutation Status in Brain Gliomas. *Neuro Oncol* (2020) 22(3):402–11. doi: 10.1093/neuonc/noz199
64. Hegi ME, Diserens AC, Gorlia T, Hamou MF, De Tribolet N, Weller M, et al. MGMT Gene Silencing and Benefit From Temozolomide in Glioblastoma. *N Engl J Med* (2005) 352(10):997–1003. doi: 10.1056/NEJMoa043331
65. Kaina B, Christmann M, Naumann S, Roos WP. MGMT: Key Node in the Battle Against Genotoxicity, Carcinogenicity and Apoptosis Induced by Alkylating Agents. *DNA Repair (Amst)* (2007) 6(8):1079–99. doi: 10.1016/j.dnarep.2007.03.008
66. Qian XC, Brent TP. Methylation Hot Spots in the 5' Flanking Region Denote Silencing of the O6-methylguanine-DNA Methyltransferase Gene. *Cancer Res* (1997) 57(17):3672–7.
67. Mansouri A, Hachem LD, Mansouri S, Nassiri F, Laperriere NJ, Xia D, et al. MGMT Promoter Methylation Status Testing to Guide Therapy for Glioblastoma: Refining the Approach Based on Emerging Evidence and Current Challenges. *Neuro Oncol* (2019) 21(2):167–78. doi: 10.1093/neuonc/noy132
68. Kazerooni AF, Bakas S, Saligheh Rad H, Davatzikos C. Imaging Signatures of Glioblastoma Molecular Characteristics: a Radiogenomics Review. *J Magnetic Resonance Imaging* (2019) 52(1):54–69. doi: 10.1002/jmri.26907
69. Lohmann P, Galldiks N, Kocher M, Heinzel A, Filss CP, Stegmayr C, et al. Radiomics in Neuro-Oncology: Basics, Workflow, and Applications. *Methods* (2020) 188:112–21. doi: 10.1016/j.ymeth.2020.06.003
70. Lotan E, Jain R, Razavian N, Fatterpekar GM, Lui YW. State of the Art: Machine Learning Applications in Glioma Imaging. *AJR Am J Roentgenol* (2019) 212(1):26–37. doi: 10.2214/AJR.18.20218
71. Xi YB, Guo F, Xu ZL, Li C, Wei W, Tian P, et al. Radiomics Signature: a Potential Biomarker for the Prediction of MGMT Promoter Methylation in Glioblastoma. *J Magn Reson Imaging* (2018) 47(5):1380–7. doi: 10.1002/jmri.25860
72. Sasaki T, Kinoshita M. Radiomics and MGMT Promoter Methylation for Prognostication of Newly Diagnosed Glioblastoma. *Sci Rep* (2019) 9: (1):14435. doi: 10.1038/s41598-019-50849-y
73. Kong Z, Lin Y, Jiang C, Li L, Liu Z, Wang Y, et al. (18)F-FDG-PET-Based Radiomics Signature Predicts MGMT Promoter Methylation Status in Primary Diffuse Glioma. *Cancer Imaging* (2019) 19(1):58. doi: 10.1186/s40644-019-0246-0
74. Chen X, Zeng M, Tong Y, Zhang T, Fu Y, Li H, et al. Automatic Prediction of MGMT Status in Glioblastoma Via Deep Learning-Based MR Image Analysis. *BioMed Res Int* (2020) 2020:9258649. doi: 10.1155/2020/9258649
75. Korfiatis P, Kline TL, Lachance DH, Parney IF, Buckner JC, Erickson BJ. Residual Deep Convolutional Neural Network Predicts MGMT Methylation Status. *J Digit Imaging* (2017) 30(5):622–8. doi: 10.1007/s10278-017-0009-z
76. Heimberger AB, Hlatky R, Suki D, Yang D, Weinberg J, Gilbert M, et al. Prognostic Effect of Epidermal Growth Factor Receptor and EGFRvIII in Glioblastoma Multiforme Patients. *Clin Cancer Res* (2005) 11(4):1462–6. doi: 10.1158/1078-0432.ccr-04-1737
77. Louis DN, Giannini C, Capper D, Paulus W, Figarella-Branger D, Lopes MB, et al. cIMPACT-NOW Update 2: Diagnostic Clarifications for Diffuse Midline Glioma, H3 K27M-Mutant and Diffuse Astrocytoma/Anaplastic Astrocytoma, IDH-Mutant. *Acta Neuropathol* (2018) 135(4):639–42. doi: 10.1007/s00401-018-1826-y
78. Aboian MS, Solomon DA, Felton E, Mabray MC, Villanueva-Meyer JE, Mueller S, et al. Imaging Characteristics of Pediatric Diffuse Midline Gliomas With Histone H3 K27M Mutation. *AJNR Am J Neuroradiol* (2017) 38(4):795–800. doi: 10.3174/ajnr.A5076
79. Liu TT, Achrol AS, Mitchell LA, Rodriguez SA, Feroze A, Iv M, et al. Magnetic Resonance Perfusion Image Features Uncover an Angiogenic Subgroup of Glioblastoma Patients With Poor Survival and Better Response to Antiangiogenic Treatment. *Neuro-oncology* (2017) 19(7):997–1007. doi: 10.1093/neuonc/now270
80. Colen RR, Wang J, Singh SK, Gutman DA, Zinn PO. Glioblastoma: Imaging Genomic Mapping Reveals Sex-Specific Oncogenic Associations of Cell Death. *Radiology* (2015) 275(1):215–27. doi: 10.1148/radiol.14141800
81. Lee JK, Wang J, Sa JK, Ladewig E, Lee HO, Lee IH, et al. Spatiotemporal Genomic Architecture Informs Precision Oncology in Glioblastoma. *Nat Genet* (2017) 49(4):594–9. doi: 10.1038/ng.3806
82. Hu LS, Hawkins-Daarud A, Wang L, Li J, Swanson KR. Imaging of Intratumoral Heterogeneity in High-Grade Glioma. *Cancer Lett* (2020) 477:97–106. doi: 10.1016/j.canlet.2020.02.025
83. Cho HR, Jeon H, Park CK, Park SH, Choi SH. Radiogenomics Profiling for Glioblastoma-Related Immune Cells Reveals CD49d Expression Correlation

With MRI Parameters and Prognosis. *Sci Rep* (2018) 8(1):16022. doi: 10.1038/s41598-018-34242-9

84. Lin P, Peng YT, Gao RZ, Wei Y, Li XJ, Huang SN, et al. Radiomic Profiles in Diffuse Glioma Reveal Distinct Subtypes With Prognostic Value. *J Cancer Res Clin Oncol* (2020) 146(5):1253–62. doi: 10.1007/s00432-020-03153-6
85. Wu J, Mayer AT, Li R. Integrated Imaging and Molecular Analysis to Decipher Tumor Microenvironment in the Era of Immunotherapy. *Semin Cancer Biol* (2020) 20(S1044-579X):30264–9. doi: 10.1016/j.semcancer.2020.12.005
86. Aslan K, Turco V, Blobner J, Sonner JK, Liuzzi AR, Núñez, et al. Heterogeneity of Response to Immune Checkpoint Blockade in Hypermutated Experimental Gliomas. *Nat Commun* (2020) 11(1):931. doi: 10.1038/s41467-020-14642-0
87. Booth TC, Williams M, Luis A, Cardoso J, Ashkan K, Shuaib H. Machine Learning and Glioma Imaging Biomarkers. *Clin Radiol* (2020) 75(1):20–32. doi: 10.1016/j.crad.2019.07.001
88. Liu X, Li Y, Sun Z, Li S, Wang K, Fan X, et al. Molecular Profiles of Tumor Contrast Enhancement: a Radiogenomic Analysis in Anaplastic Gliomas. *Cancer Med* (2018) 7(9):4273–83. doi: 10.1002/cam4.1672
89. Grossmann P, Narayan V, Chang K, Rahman R, Abrey L, Reardon DA, et al. Quantitative Imaging Biomarkers for Risk Stratification of Patients With Recurrent Glioblastoma Treated With Bevacizumab. *Neuro Oncol* (2017) 19 (12):1688–97. doi: 10.1093/neuonc/nox092
90. Petrova L, Korfiatis P, Petr O, Lachance DH, Parney I, Buckner JC, et al. Cerebral Blood Volume and Apparent Diffusion Coefficient - Valuable Predictors of non-Response to Bevacizumab Treatment in Patients With Recurrent Glioblastoma. *J Neurol Sci* (2019) 405:116433. doi: 10.1016/j.jns.2019.116433
91. Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, et al. The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. *J Digit Imaging* (2013) 26(6):1045–57. doi: 10.1007/s10278-013-9622-7
92. Edwards NJ, Oberti M, Thangudu RR, Cai S, Mcgarvey PB, Jacob S, et al. The CPTAC Data Portal: A Resource for Cancer Proteomics Research. *J Proteome Res* (2015) 14(6):2707–13. doi: 10.1021/pr501254j
93. Wang LB, Karpova A, Gritsenko MA, Kyle JE, Cao S, Li Y, et al. Proteogenomic and Metabolomic Characterization of Human Glioblastoma. *Cancer Cell* (2021) 39(4):509–528 e20. doi: 10.1016/j.ccell.2021.01.006
94. Puchalski RB, Shah N, Miller J, Dalley R, Nomura SR, Yoon JG, et al. An Anatomic Transcriptional Atlas of Human Glioblastoma. *Science* (2018) 360 (6389):660–3. doi: 10.1126/science.aaf2666
95. Zwanenburg A, Vallieres M, Abdalah MA, Aerts H, Andrearczyk V, Apte A, et al. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-Based Phenotyping. *Radiology* (2020) 295(2):328–38. doi: 10.1148/radiol.2020191145
96. Shao L, Liu Z, Feng L, Lou X, Li Z, Zhang XY, et al. Multiparametric MRI and Whole Slide Image-Based Pretreatment Prediction of Pathological Response to Neoadjuvant Chemoradiotherapy in Rectal Cancer: a Multicenter Radiopathomic Study. *Ann Surg Oncol* (2020) 27(11):4296–306. doi: 10.1245/s10434-020-08659-4

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Liu, Chen, Hu, Yang, Liu, Hu, Ge, Zhang and Liu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/8683403

Functional Genomics Approaches in Arthritis

Article *in* American Journal of PharmacoGenomics · February 2004

DOI: 10.2165/00129785-200404010-00004 · Source: PubMed

CITATIONS
34

READS
71

3 authors, including:

Mukundan Attur
NYU Langone Medical Center
182 PUBLICATIONS **10,651** CITATIONS
SEE PROFILE

Mandar Dave
New York University
56 PUBLICATIONS **1,734** CITATIONS
SEE PROFILE

All content following this page was uploaded by Mukundan Attur on 30 March 2022.

The user has requested enhancement of the downloaded file.

GENOMICS IN HEALTH AND DISEASE

Am J Pharmacogenomics 2004; 4 (1): 29-43
1175-2203/04/0001-0029/$31.00/0

© 2004 Adis Data Information BV. All rights reserved.

Functional Genomics Approaches in Arthritis

Mukundan G. Attur,[1] *Mandar N. Dave*[1] and *Ashok R. Amin*[1,2,3]

1 Departments of Rheumatology and Medicine, Hospital for Joint Diseases, New York, New York, USA
2 Departments of Pathology & Medicine, New York University School of Medicine, New York, New York, USA
3 Kaplan Cancer Center, New York, New York, USA

Contents

Abstract ... 29
1. Genetic Complexity and Arthritis ... 30
2. Gene Mining in Arthritis ... 30
  2.1 Gene Expression Analyses ... 30
  2.2 Genome-Wide Scans, Association Studies, and Positional Cloning ... 30
3. Validating Target Hypothesis ... 31
4. Functional Genomics ... 32
  4.1 Cell Culture Systems ... 32
  4.2 Functional Genomics in Animal Models ... 33
    4.2.1 Development of Autoimmune Transgenic Mice for Rheumatoid Arthritis ... 34
    4.2.2 An *In Vivo* Model of SCID Mouse for Human Cartilage Invasion ... 34
    4.2.3 Genetic Disorders in Human Cartilage/Skeleton ... 34
    4.2.4 Other Animal Models ... 34
  4.3 Molecular Targets in Arthritis ... 35
    4.3.1 Matrix Metalloproteases (MMPs) ... 35
    4.3.2 Extracellular Matrix Components ... 36
    4.3.3 Cytokines ... 38
5. Conclusions and Future Directions ... 40

Abstract

The post-genomic era of functional genomics and target validation will allow us to narrow the bridge between clinically correlative data and causative data for complex diseases, such as arthritis, for which the etiological agent remains elusive. The availability of human and other annotated genome sequences, and parallel developments of new technologies that allow analysis of minute amounts of human and animal cells (peripheral blood cells and infiltrating cells) and tissues (synovium and cartilage) under different pathophysiological conditions, has facilitated high-throughput gene mining approaches that can generate vast amounts of clinically correlative data. Characterizing some of the correlative/causative genes will require reverting to the hypothesis-driven, low throughput method of complementary experimental biology using genomic approaches as a tool. This will include *in silico* gene expression arrays, genome-wide scans, comparative genomics using various animal models (such as rodents and zebrafish), bioinformatics and a team of well trained translational scientists and physicians.

For the first time, the 'genomic tools' will allow us to analyze small amounts of surgical samples (such as needle biopsies) and clinical samples in the context of the whole genome. Preliminary genomic analysis in osteoarthritis has already resurrected the debate on the semantic issues in the definition of inflammation. Further analyses will not only facilitate the development of unbiased hypotheses at the molecular level, but also assist us in the identification and characterization of novel targets and disease markers for pharmacological intervention, gene therapy, and diagnosis.

1. Genetic Complexity and Arthritis

Arthritis is a complex disease with an unknown etiology. Based on the clinical symptoms, it can be classified as osteoarthritis (OA), rheumatoid arthritis (RA), synovial lipomatosis, avascular necrosis, crystal deposition disease, gout, and other minor diseases. The common underlining symptoms in the above clinical manifestations include destruction of joints and inflammation.[1]

Unlike arthritis, monogenetic diseases (e.g. cystic fibrosis) have been identified where one single mutation is responsible for the dysfunction of the gene leading to the disease. It took several years to map and characterize the gene responsible for cystic fibrosis.[2] The development of new genomic approaches and tools has facilitated the identification of over 40 new genes known to be involved in various diseases in the last 3 years,[3] which are mostly the low hanging fruits of this vast sea of information. The major challenge faced is characterization of genes involved in oligo- and polygenic disorders such as arthritis, because pedigrees from complex diseases reveal no Mendelian inheritance patterns, and single gene mutations are neither sufficient nor necessary to explain the disease phenotypes.

Many traits of either the normal or diseased state in humans are controlled by a single gene and, thus, have a distinct, discontinuous, phenotype (e.g. blood type). There are other traits where a continuous distribution is observed in a segregating population (e.g. weight gain, IQ, learning ability, etc.), yielding a bell-shaped curve when analyzed quantitatively. Thus, continuous traits are often given a quantitative value, referred to as quantitative trait. Furthermore, the genetic loci controlling these traits are called the quantitative trait loci (QTL). Mapping markers linked to QTL helps identify regions of the genome that may contain genes involved in the expression of the quantitative trait.

QTL are generally regulated by several genes.[4] This is not a rule of thumb because in non-animal species positional cloning of a QTL revealed that a single gene was responsible for the complex regulation of tomato fruit size.[5] It is possible that some of the human QTL may also prove to have a single gene responsible for complex disorders. A preliminary map of QTL in mice has recently been published, which may facilitate structure-functional analysis of similar genes in humans.[6-9] Human arthritis may involve several genes and gene products influenced by physical, environmental, and other risk factors. Furthermore, QTL have been identified for arthritis in rat models (e.g. *Cia5*, *Cia7*, *Aia3/Cia3*, etc.).[10-13] The possibility of QTL being involved in human arthritis cannot be discounted.

Advances in various molecular biology tools and information in genomics now permit the detailed study of complex diseases such as arthritis, for example, the etiology and the pathogenesis of joint destruction. The greatly enhanced flow of information provides us with an opportunity to reassess our working hypothesis. Such a reassessment allows us to unravel the physical or biological basis of the mysteries of complex diseases.

2. Gene Mining in Arthritis

Basically, there are two approaches to gene mining in complex diseases that have been used in arthritis and inflammatory diseases. The first approach is to compare gene expression arrays of normal and diseased target tissues. The second approach for gene mining is using genome-wide scans, association studies and positional cloning analysis utilizing microsatellite marker loci, sibling pairs and disease-susceptible families.

2.1 Gene Expression Analyses

Good examples of target tissues for analysis of differential gene expression in arthritis are cartilage, synovium, and blood leukocytes. Figure 1 shows an approach using complementary technologies for human cartilage. We[14,15] and others[16-19] have applied some of these strategies. Differential gene expression using subtractive hybridization of synovium for RA and OA has also been examined.[20,21]

It is estimated that 10 000–20 000 transcripts and 50 000–80 000 proteins are expressed in a given cell at a certain time period. Therefore, the gene mining technologies described in figure 1, either individually or in combination, have the tendency to generate an enormous number of differentially expressed transcripts, expressed sequence tags (ESTs) and proteins, which may be involved (either directly or indirectly) in the disease process. These potential targets have to be reduced to a manageable number by prioritizing the ones that may be involved in the disease process. Prioritization can be based on criteria such as tissue and/or cell specificity, and relative abundance (see Attur et al.[23] for a recent review). Computational data analysis and clustering algorithms can be optimized for each project to extract relevant information. Another strategy to reduce targets to a manageable number would be to apply pharmacogenomic screens (*in vitro* and *in vivo*), and to compare the targets to chemogenomic databases during pharmacological intervention using various reagents/drugs such as anti-tumor necrosis factor-α (TNFα) and/or recombinant TNF antagonists.[24]

2.2 Genome-Wide Scans, Association Studies, and Positional Cloning

Genome-wide scans provide the means to identify several trait-related genetic factors. In this approach, a large collection of genetic markers is tested for association with the disease. These markers are distributed frequently enough across the chromosomes so as to cover the whole genome. By themselves, they may or may not be related to the diseased state, however, their proximity to a susceptibility gene may reflect their effect on the gene.

© 2004 Adis Data Information BV. All rights reserved.

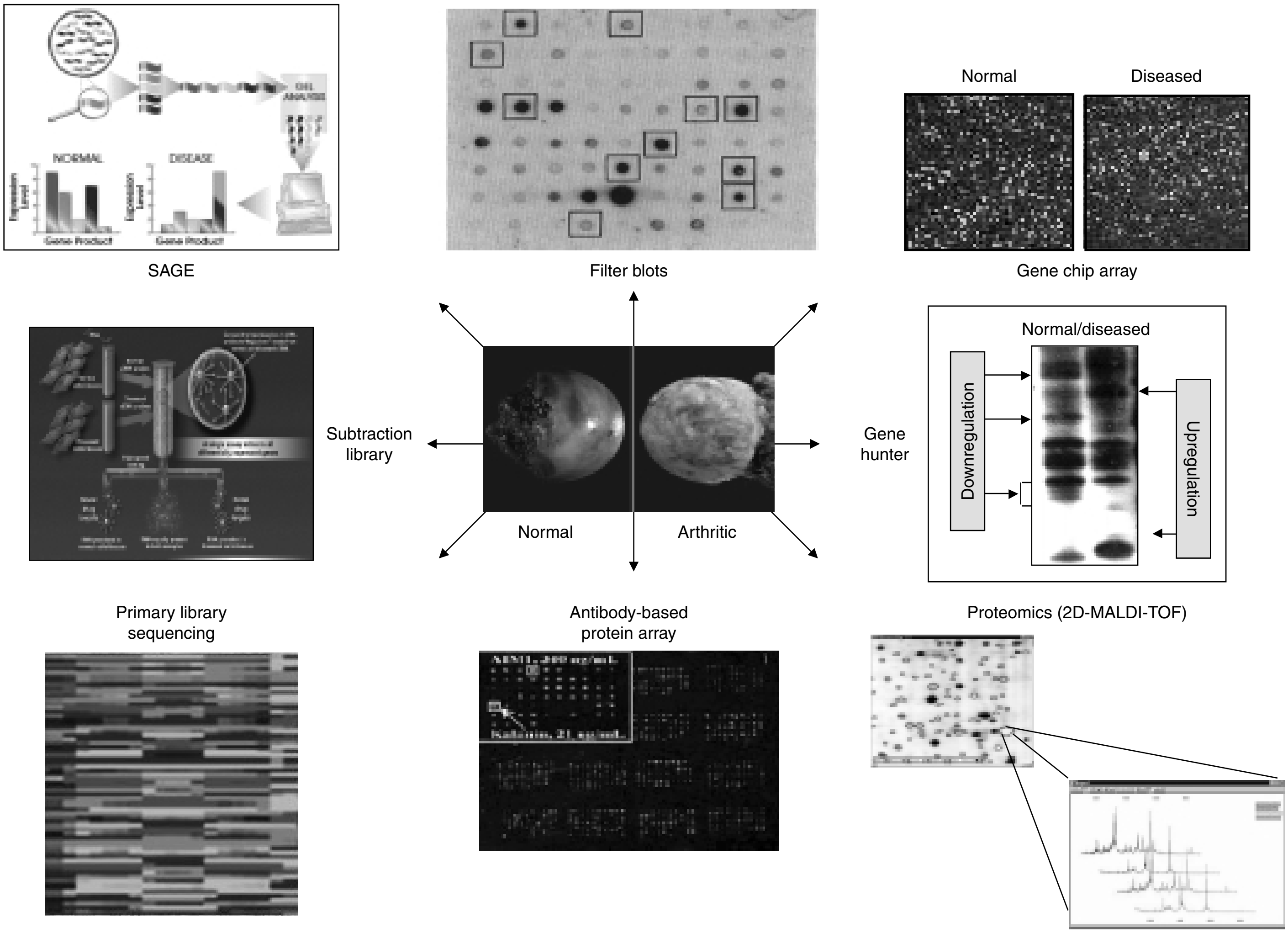


Fig. 1. Gene mining in human cartilage. There are several methods of gene mining in cartilage and differential expression of genes/proteins in normal and arthritis-affected cartilage. Total RNA and protein can be extracted from human cartilage and subjected to various analyses. These include oligonucleotide-based gene chip arrays on glass and filter membranes, differential display of arbitrary nested primers (Gene Hunter Kits, Gene Hunter Corp., Nashville, TN, USA), proteomics with two-dimensional matrix-assisted laser desorption time-of-flight (2D-MALDI-TOF), antibody based protein arrays, high-throughput DNA sequencing of primary cDNA libraries and substraction libraries, preparation of cDNA libraries for Serial Analysis of Gene Expression (SAGE) and radiolabeled dot (filter) blots. These technologies on the whole, cover a broad area of gene mining and complement each other.[22]

Examination of the outcomes of such multiple testing processes may potentially uncover regions in the genome containing a gene (or genes) associated with the trait.

Positional cloning of genes based on their position on the genomic map is a method of choice for isolating genes for which only a mutant phenotype is known. Many genetic disorders result from mutations in genes whose phenotype (or protein product) and/or function is unknown. Several genes have been identified using this approach, one of the earlier ones being the X-linked gene for chronic granulomatous disease in 1996. Other examples include the *IBD5* gene for Crohn's disease, and genes associated with Huntington disease, cystic fibrosis, and obesity.[25]

Recently, large cohort genotypic studies suggested that the susceptibility genes for human OA and RA have been mapped on chromosome 16p between 28cM and 47cM from the telomere at 1(D1S235), 4(D4S1647), 12(D12S373), 16(D16S403 and D16S401), 17(D17S1301), 10(D10S2327 and D10S201), 14(D14S587 and D14S285), which overlap with other autoimmune and inflammatory diseases.[26,27]

3. Validating Target Hypothesis

Target validation serves as an important bridge between correlative data and causative data. Figure 2 summarizes the methods to generate hypothesis-driven correlative data and validation of these against causative data sets. The ultimate biological validation of a pharmaceutical target comes with the successful phase III clinical trial and broad application of the therapy to a large population of patients.

© 2004 Adis Data Information BV. All rights reserved.

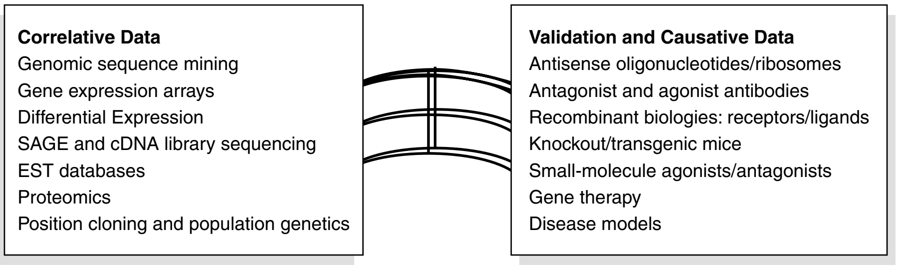

Correlative Data
Genomic sequence mining
Gene expression arrays
Differential Expression
SAGE and cDNA library sequencing
EST databases
Proteomics
Position cloning and population genetics
Validation and Causative Data
Antisense oligonucleotides/ribosomes
Antagonist and agonist antibodies
Recombinant biologies: receptors/ligands
Knockout/transgenic mice
Small-molecule agonists/antagonists
Gene therapy
Disease models

Fig. 2. Approaches to bridge correlative data with validation and causative data. There are a number of methods to accumulate correlative data, as shown in figure 1. Potential targets developed using various screening methods require a specialized approach to validate or exclude them as causative targets before they can be ultimately validated in humans through clinical trials. Bioinformatics plays an essential role in bridging the two areas. **EST** = expressed sequence tag; **SAGE** = Serial Analysis of Gene Expression.

Here we focus on correlating the target hypothesis to the disease under study using various examples in arthritis. Validation procedures have become an important component of gene mining studies because of the following difficulties: (i) much of these gene mining technologies have not been compared in parallel with one another using the same clinical samples; (ii) some of the commercially available gene chips had serious flaws in probe design and reproducibility, as evidenced recently;[28] (iii) the new generation of commercial chips (with increasing probe sets such as U133A from Affymetrix Corporation, CA, USA) and the updated microarray data analysis software are not compatible with the previous generation of chips (e.g. U95); (iv) some of the low abundant and short half-life transcripts (e.g. membrane proteins and some cytokines) are not detected in clinical samples when using these gene chip arrays; and (v) variation in the read-outs (^{32}P or fluorescent labeled probes, pseudo-colors or semi-quantitative methods). All of these shortfalls impede cross-sectional statistical analysis for target validation. There are some reports which address these issues.[29-32]

Real-time PCR has been a method of choice for validating differentially expressed transcripts.[33,34] We have successfully identified and validated low abundance transcripts (5–100 copies of mRNA) involved in inflammation that were missed in gene chip arrays.[34] Subcellular localization of differentially-expressed targets by *in situ* staining using antibodies or riboprobes is important. This has recently been demonstrated using differentially expressed target genes such as *SPP1* (secreted phosphoprotein-1 or osteopontin), *EGR1* (early growth response-1), and matrix metalloproteinase genes *MMP1*, *-3*, *-8*, *-9* and *-13* in arthritis.[35,36] Gene expression patterns of *MMP1*, *MMP3* and *MMP9* in OA-affected cartilage were found to be zonal and grade-specific.[37] These approaches can allow validation of targets across sections of clinical samples.

4. Functional Genomics

Functional genomic analysis involves a systematic effort to understand the function of genes and gene products and biological systems classically performed for single genes in the context of the whole genome. Functional genomics can be conceptually divided into two approaches:

1. the gene-driven approach, where one uses genomic information for identifying, cloning, expressing, and characterizing the gene at the molecular level
2. the phenotype-driven approach, which analyzes phenotypes from random mutation screens or naturally occurring variants (mouse mutants, human diseases) to identify and characterize genes associated with the phenotype, without prior knowledge of the underlying molecular mechanism or function.

These strategies are complimentary, leading collectively to association of phenotypes with genotypes. As functional genomics begins to mature into a coherent science (as molecular biology did in the last half of last century) its constituent fields become clearer. As shown in figure 3, they include bioinformatics, structural genomics, comparative genomics, expression genomics, and proteomics. The use of these fields in studying arthritis, with emphasis on genes encoding MMPs, matrix proteins/proteoglycans, cytokines and cytokine receptors, are discussed in section 4.3.

4.1 Cell Culture Systems

Several different cell culture systems have also been used for functional analysis in cartilage and chondrocytes. Primary chondrocytes (outside the cartilage) have a tendency to dedifferentiate in *in vitro* systems. However, recent studies by Goldring et al.,[38] have shown immortalization of human chondrocytes using retroviral vectors expressing SV40-large T antigen, to maintain the chondrocyte phenotype without dedifferentiating into fibroblast-like cells. These cells, although they lack several important phenotypic characters of a chondrocyte such as the production of nitric oxide (NO) and prostaglandin E_2 (PGE_2), have proven to be useful for studying a variety of general and chondrocyte specific functions.

An approach to restore the differentiated properties of chondrocytes is by cultivating them in alginate beads, alginate-fibrin carrier, or other scaffolds for cartilage/chondrocyte culture sys-

© 2004 Adis Data Information BV. All rights reserved.

tems.[39-41] This system supports cell proliferation without the loss of functional potential. Furthermore, the cells can be retrieved for analysis by dissolving monomer subunits using chelating agents. Another approach is to grow primary chondrocytes in collagen II coated dishes, allowing the chondrocytes to wrap around their own matrix and maintaining the chondrocyte-like phenotype.[42] Alternatively, chondrosarcoma cells, which retain much of the chondrocyte phenotype, can also be used for functional analysis.[43] Immortalized human synovial cells have also been used for similar studies.[44] However, the choice of chondrocyte culture system for *in vitro* studies for pharmacological intervention should be chosen keeping in mind the expected goals or questions asked.

4.2 Functional Genomics in Animal Models

Animal models of arthritis are somewhat self-limiting because they aim to reproduce a human disease phenotype that is heterogeneous in nature. There are no clear-cut genetic models of arthritis (OA or RA) as the etiological gene(s) under specific genetic background are elusive. At least three examples can be cited for this.

1. Mutations in camptodactyly-arthropathy-coxa vara-pericarditis (CACP) or lubricin or proteoglycan 4 (PRG4) in humans show flexion contractures of all fingers and bilateral swelling to destruction of joints.[45] Marcelino et al.[46] developed homozygous knockout mice for CACP. This mouse showed hyperplasia of synviocytes and increased matrix deposition in the subintimal layer. These studies show similar phenotypic characteristics in humans and mice with similar genotypioc defects.[45]
2. Mutation in *COL10A1* leads to Schmid metaphyseal chondrodysplasia (SMCD) in humans. Recent studies by Chan et al.[47] in mice showed mutation in a single allele of *COL10A1* (haploinsufficiency) also demonstrated SMCD similar to humans. These studies suggest the importance of gene dosage *per se* in animal models of arthritis.
3. Mutation in the MMP2 gene in humans causes multicentric osteolysis and arthritis syndrome[48] whereas the MMP2 null mice

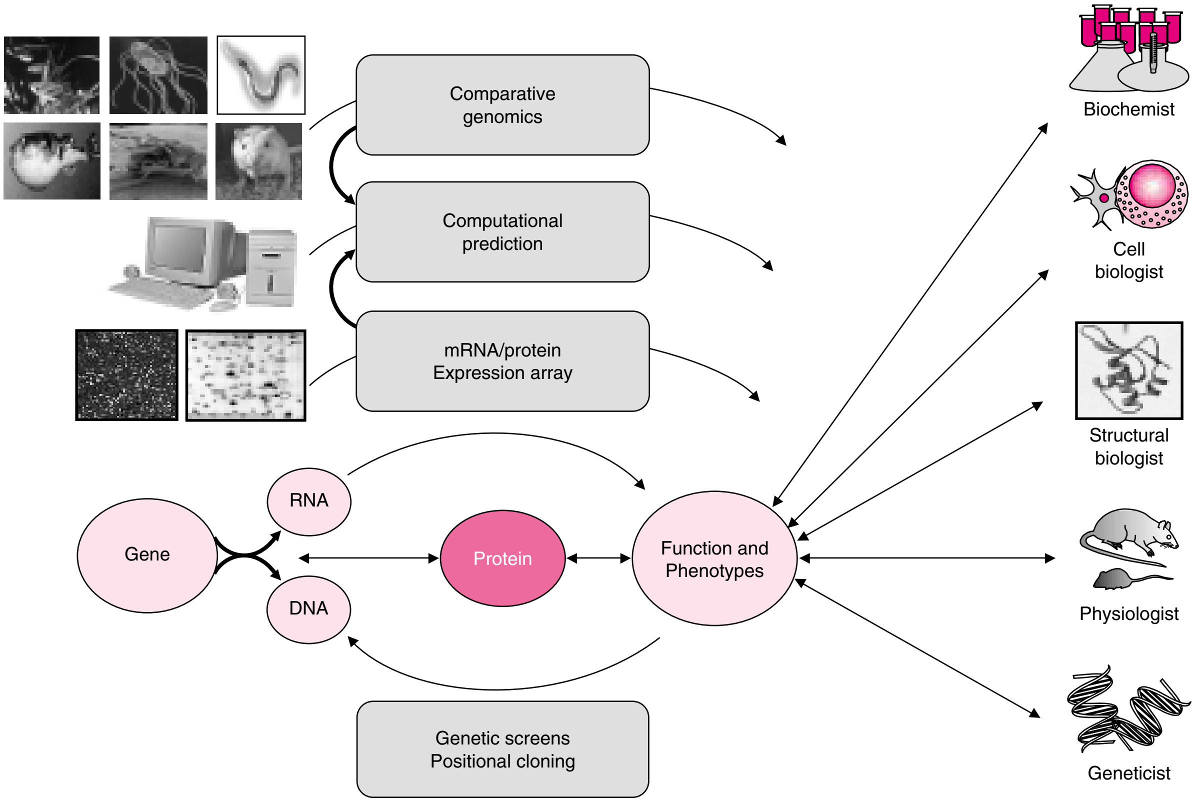


Fig. 3. A matrix approach to functional genomics. A global multidimensional integrated program is required with respect to tools, technology, and training of individuals in various areas of biological sciences to fully exploit the power of functional genomics in medicine. The present gene-mining approaches can identify target genes and/or biomarkers outside the area of expertise of the investigator and, hence, require adaptation for crossing over to other fields. Gene expression arrays and proteomics remain important tools to develop working hypotheses and also validating targets. Computational predictions facilitate hypothesis-driven prediction, clustering analysis and preparation of databases for future references. Bioinformatics also bridges analysis between comparative genomics and expression arrays (mRNA/protein) for functional genomics. Comparative genomics using relatively simple genome structure of microbes, nematodes, arthropods, Zebrafish, puffer fish, mice, and other lower vertebrates, becomes a key in dissecting and validating targets for complex human diseases. These strategies bridge and complement the gene-driven approaches and phenotype-driven approaches in diseased individuals, animal models or dysfunctional cell types to identify and validate targets.

© 2004 Adis Data Information BV. All rights reserved.

developed normally, except that there is a subtle delay in their growth. However, MT1-MMP (membrane type MMP1; MMP14) null mice showed a phenotype similar to osteolysis patients even though both MMPs are implicated in the pathophysiological process in arthritis in humans. It should be noted that MMP14 is a regulator of MMP2. These studies suggest that similar phenotypes are observed in humans and mice with closely regulated or related MMPs.

Therefore, multiple approaches may be necessary for target validation. Furthermore, genes that have not previously been implicated in OA, have been identified using this approach.[23] There are no fixed criteria to select a target for such mouse models and it all depends on the question one is asking.

4.2.1 Development of Autoimmune Transgenic Mice for Rheumatoid Arthritis

Joint degradation characteristic of RA is a result of an autoimmune response. Various genomic technologies have facilitated the identification of novel auto-antigens that may play a role in arthritis. These can be functionally examined by using transgenic animals. A transgenic mouse (K/BxN T cell receptor [TCR]) spontaneously develops a joint disorder with many of the clinical, histological, and immunological features of human RA.[49] The disease is initiated by a subset of T cells recognizing a self antigen identified as glucose-6-phosphate isomerase, a glycolytic enzyme. Thus some forms of RA or related arthritis may be developed by a mechanism fundamentally different from the current paradigm of joint-specific T cell response.

4.2.2 An In Vivo Model of SCID Mouse for Human Cartilage Invasion

Synovial hypertrophy and pannus formation play a critical role in inflammation and cartilage destruction in RA. In view of this, Jorgensen and Gay[50] have developed a human synovial fibroblast/cartilage interaction model *in vivo* to examine the role of various genes. This model allows examination of synovial cell activation, attachment to cartilage and its role (and mechanism) in destruction of cartilage. Briefly, human RA-affected synovial fibroblasts are grown *in vitro* and transfected with a gene of choice using a retroviral vector. The cells are then packaged in an inert sponge together with normal human cartilage and implanted in a renal capsule in severe-combined immunodeficient (SCID) mice, which do not evoke an immune response to the transplantation. The gene expression is mostly stable for 60–80 days, during which time various forms of bioassays and analyses can be performed.[50] This strategy has been utilized to examine the effect of interleukin (IL)-1 receptor antagonist (IL-1Ra), TNF receptor I (TNFRI), IL-10, tumor suppressors PTEN (phosphatase and tensin homolog) and p53 in cartilage degradation.[24,51,52]

4.2.3 Genetic Disorders in Human Cartilage/Skeleton

The cartilage extracellular matrix is composed of numerous macromolecules. The major constituent is collagen fibrils composed mainly of type II collagen which in turn interacts with various proteins, which include peptidoglycan and aggrecans. Recently several studies have shown that loss of cartilage matrix protein such as collagen has drastic effect on skeletal development. The mouse mutant for both extracellular matrix (ECMs) proteins showed severe and lethal chondrodysplasia.[53-55] Recently Mier et al.[56] have also shown that mutations in type II precollagen gene in children results in precocious osteoarthritis. Conversely, Baldwin et al.[57] have shown that there is no linkage or association between type II collagen mutation and osteoarthritis.

Chondrocytes express collagen type II, V, VI, IX, X, XII, and XIV, depending on their physiological stage.[58] Approximately 278 different mutations have been reported to date in genes for type I, II, III, IX, X, and XI collagens in various diseases which impact bone, cartilage, and blood vessels. A majority (78%) of the mutations are single-base changes either alter the codon of a critical amino acid or lead to abnormal RNA splicing, which may lead to a spectrum of diseases of bone and cartilage including osteogenesis imperfecta (a type of chondrodysplasia) and OA.[59] Some of the phenotypes of these mutations have been examined in mouse models.

Collagen type IX is a non-fibrillar collagen localized on the surface of type II collagen. Alternatively spliced forms of collagen type IX are expressed on hyaline cartilage and other tissues. A mouse strain lacking both forms showed no detectable abnormalities at birth, but developed a severe degenerative joint disease resembling human OA.[60] Similar experiments by Nakata et al.[61] using a tissue-specific promoter/enhancer to express type IX collagen revealed pathological changes similar to human OA and chondrodysplasia.

The interaction between different collagens plays an important role in cartilage homeostasis. Although mice deficient in collagen II (Col2a1−/−) die at birth, and Col9a1−/−develop OA-like phenotypes, the Col2a1+/−Col9a1−/−mice show no accelerated OA.[62] These observations suggest that extracellular matrix proteins have different roles during cartilage development. Type II collagen may be important for cartilage formation, and type IX collagen may be required for cartilage maintenance.

4.2.4 Other Animal Models

Zebrafish serve as a powerful experimental model for functional genomics in identification of developmentally important genes that can be related to human development and disease.[63] In zebrafish, the cartilage of the pharynx develops during late embryogenesis and grows extensively in the larva before eventually being replaced by bone. The detectable expression of genes both during development as well as in the adult, and the potential for studying gene interactions and over-expression all suggest that this

© 2004 Adis Data Information BV. All rights reserved.

model system would contribute substantially to the understanding of this complex system of cartilage and bone development.

4.3 Molecular Targets in Arthritis

4.3.1 Matrix Metalloproteases (MMPs)

Matrix metalloproteases, also called matrixins, play major roles in extracellular matrix turnover. Dysfunctional expression of MMP activity constitutes part of the pathogenic mechanism associated with a wide range of diseases. These include tissue breakdown and remodeling during invasive tumor growth and tumor angiogenesis,[64] degradation of myelin-basic protein in neuroinflammatory diseases,[65] blood-brain barrier following brain injury,[66] increased matrix turnover in atherosclerotic plaques,[67] loss of aortic wall strength in aneurysms,[68] tissue degradation in gastric ulceration,[69] and liver fibrosis.[70] MMP expression has been extensively studied in OA and RA. The expression of MMPs (MMP-1, -3, -9 and -13) in arthritis is strongly stimulated by inflammatory cytokines such as IL-1 and tumor necrosis factor-α (TNFα).[37,71]

MMPs belong to a superfamily of zinc-dependent proteases known as metzincins. Based on sequence and structural similarities, metzincins are grouped into four distinct subfamilies: the astacins, matrixins (MMPs), the adamalysins (or snake venom proteases and ADAMs), and the seralysins (large bacterial proteinases). The term ADAMs stands for *a d*isintegrin and *m*etalloprotease, which represent the two key structural domains in these molecules. ADAMs are distinct among cell surface proteins in containing features of both adhesive proteins and proteinases, and their roles in cell-cell interactions have attracted particular interest in recent times.

Recently, new members of the ADAM family, known as ADAMTS (*ad*isintegrin *a*nd *m*etalloprotease with thrombospondin motifs), have been characterized, which include aggreganase (ADAMTS-4 and -5).[72] The ADAMTS are also distinguished from the ADAMs by their lack of cysteine rich, endothelial growth factor (EGF)-like, transmembrane, and cytoplasmic domains. The structure/function relationships of various MMPs (ADAMs and ADAMTS) are elegantly reviewed elsewhere.[73,74] MMP activity is also regulated by endogenous inhibitory activity, attributed to a family of four tissue inhibitors of metalloproteases (TIMPs). TIMPs act by forming a 1 : 1 complex with the activated catalytic zinc-binding domain in the MMPs. Chronic stimulation of MMP activities due to an imbalance in the levels of MMPs and TIMPs has also been implicated in the pathogenesis of variety of diseases such as cancer, OA and RA.[75]

Among the 25 vertebrate MMPs identified to date, 22 human homologs have been confirmed.[76,77] Each of these MMPs has distinct as well as overlapping substrate specificities, and can virtually cleave most ECM proteins involved in joint destruction in RA and OA. These enzymes are predominantly produced by macrophages and neutrophils in the synovial pannus in RA, and cartilage in OA.

Differential Expression and Identification of Proteases from Human Arthritis-Affected Cartilage

Adamalysin (ADAM) Proteases

The analysis of the human genome has allowed us to predict the presence of approximately 500 MMP-like transcripts in humans, which have yet to be functionally characterized.[78,79] Human arthritis-affected cartilage and synovium are one of the richest sources of differentially expressed disease-specific proteases.[14,36,37] We, and others, have therefore selected healthy and diseased cartilage to identify novel proteases. Gene mining efforts using total RNA from healthy, RA- and OA-affected cartilage,[14] yielded a clone (clone 8) that was up-regulated and showed two important characteristics in its partial cDNA sequence. The first is a partial cysteine switch sequence (PKVGY) and zinc binding region (HELGHN), revealed by comparative genomics as being similar to that seen in snake venom proteases.[80] Secondly, the distance between the two motifs was found to be approximately 690 base pairs. This was classified as a unique protease because these characteristics excluded the matrixin family of MMPs.

Since the sequence homology was not sufficient to predict the nature of the clone, comparative genomics and bioinformatic approaches were further combined to compare the 3-D ribbon structures of clone 8 to the catalytic region of other MMPs across species, including snake venom proteases. In spite of the low sequence homology, the three hydrophobic side chains that support 3-D folds were conserved in snake venom protease and clone 8 (figure 4), suggesting that clone 8 peptidase was structurally and functionally similar to M12B snake venom protease family. The full-length cDNA sequence of these cartilage snake venom proteases (atrolysin [C/D] and adamalysin) [figure 4] showed approximately 99% homology to TNFα convertase (TACE).[14] Although the overall sequence identity was not very high, the alignment score (20.4 for liag, which represents the protein database code for a three-dimensional structure of adamylasin II, solved at 2.0A resolution) suggested similarity between sequences of adamalysins and the clone 8[81] that was several sigmas above the noise level.

TACE is a potential target for pharmacological intervention of TNFα production, based on its ability to cleave membrane-bound proTNFα from soluble TNFα. Other putative substrates for TACE include L-selectin, TNFα receptors I and II, amyloid precursor protein (APP), IL-1 receptor and IL-6 receptor.[82]

How does TACE distinguish between all of these substrates? Preliminary data suggests that different domains of TACE are necessary for turnover of different substrates.[82] Inhibitors of TACE block TNFα activity in arthritis-affected cartilage.[83] These experiments demonstrated a functional paracrine/autocrine role of

© 2004 Adis Data Information BV. All rights reserved.

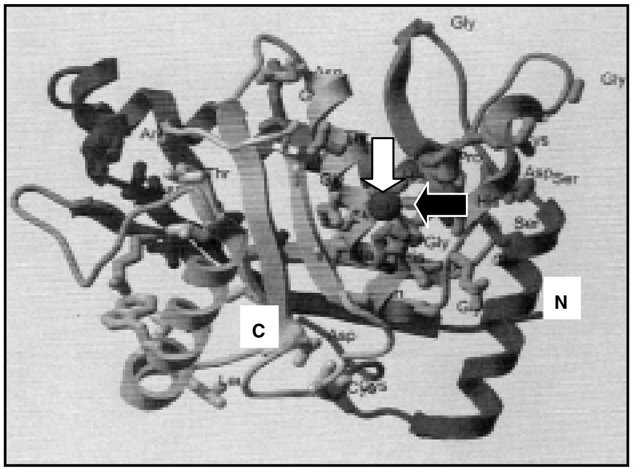


Fig. 4. Superimposition and comparison of the catalytic metallopeptidase region of atrolysin (C/D)/adamalysin with Clone 8. Human osteoarthritis (OA) and rheumatoid arthritis (RA)-affected cartilage when compared to normal cartilage showed upregulation of a partial transcript which was identified as Clone 8. The 3-dimensional structure of atrolysin (C/D)/adamalysin was used as a template. Only the residues conserved between atrolysin, adamalysin, and Clone 8 are shown by balls-and-sticks. Three conserved catalytic histidines are shown with a darkened arrow. Zinc ion is shown with an open arrow. The ribbon is marked as N to C terminus. The molecular modeling showed structural similarities in spite of low sequence homology (reproduced from Amin,[22] with permission).

TNFα in arthritis-affected cartilage that may depend, in part, on up-regulated levels of chondrocyte-derived TACE.[14] Furthermore, such *ex vivo* functional genomic assays provide real-time TNFα and TACE expression data in cartilage. Mice lacking active TACE have been generated and characterized. Although cells from TACE deficient mice release 80–90% less soluble TNFα than wild type cells, mice lacking TACE results in perinatal lethality.[84]

Serine Proteases

Hu et al.[16] utilized a known differential gene expression approach using a Gene Hunter kit (GenHunter Corp., Nashville, TN, USA), to characterize the function of a differentially expressed protease in OA-affected cartilage. This technique amplifies a family of proteases with specific domains and may exclude other important proteases. This approach picks up novel transcripts but it is labor-intensive and produces a sizable number of false positives.

Differential expression of transcripts in healthy and OA-affected human cartilage identified an evolutionarily conserved serine protease, the human high-temperature requirement A (HtrA, gene name *PRSS11*), which was up-regulated (7-fold) in human OA-affected cartilage compared with healthy cartilage.[16] Cloning and expression of *PRSS1* revealed endoproteolytic activity, including autocatalytic cleavage. The putative active site was mapped to serine 328. Recent substrate specificity studies suggest that this protease has the ability to degrade cartilage oligomeric matrix protein (COMP) and fibronectin, which are major structural proteins found in human cartilage.[85]

Identifying the Role of MMPs by a Phenotype-Driven Functional Genomic Approach

Familial osteolysis is a rare inherited disorder in which affected individuals exhibit characteristic facial features, lytic lesions of the bone, and arthritis. Linkage analysis showed *MMP2* as a candidate gene harboring a mutation in four Saudi Arabian families.[48] *MMP2*-null mice have no developmental defect, which may reflect genetic redundancy. However, it should be noted that mice with the *MMP14* gene defect have many of the same features as individuals with osteolysis and arthritis. One explanation of the difference between humans and rodent may be due to distinct regulation of *MMP2* and *MMP14*, and the balanced regulations they exert on inactive and active transforming growth factor-β (TGFβ).[86] Overexpression of TGFβ causes deregulation of bone remodeling, where as inhibition of TGFβ signaling results in cartilage destruction resembling OA.[87] This example highlights the limitation of mouse gene targeting methods despite their exquisite potential for addressing gene functions in defined contexts. This is an excellent phenotype-driven functional genomic approach wherein a defined phenotype is correlated from rodents to humans. However the drawback of this approach lies in the requirement of a pure family of familial genotype, which can be correlated to the observed phenotype.

4.3.2 Extracellular Matrix Components

Functional Genomic Analysis of Differentially Expressed Matrix Components in Cartilage

In tissues like bone and cartilage, ECM components regulate some of the most fundamental cellular processes such as growth, survival, differentiation, motility, signal transduction, and cell shape.[15,88,89] A close look at chondrocytes in cartilage by electron microscopy show that these cells are relatively larger in size than other inflammatory cells such as macrophages (~25μm vs 3.8μm in diameter and are metabolically active).[33] Furthermore, the chondrocyte cell membrane is highly invaginated (similar to microvillae) into the pericellular region of the matrix. Several types of surface receptors on chondrocytes have the ability to bind to various types of ligands in the matrix.[90] These include collagen, proteoglycans, COMP, osteopontin, and vitronectin.

Examination of the role of matrix proteins in human cartilage requires an unconventional assay system. We have developed an *ex vivo* organ culture assay to examine the role of chondrocyte/matrix interaction in cartilage without disturbing its delicate architecture (figure 5).[91] These arthritis-affected cartilage samples spontaneously release various inflammatory mediators, including NO, PGE_2, MMPs, cytokines which modulate matrix homeostasis in *ex vivo* culture.[89] Recombinant proteins such as ligands, soluble receptors, antibodies, and low molecular weight disease-modi-

© 2004 Adis Data Information BV. All rights reserved.

fying antirheumatic drugs (DMARDS), can be added to the assay to analyze inflammation and cartilage homeostasis. This approach has recently been extended into pharmaco- and chemogenomic assays for profiling transcripts in the presence of NSAIDS and DMARDS, which have the ability to modify homeostasis in cartilage.[92-94] This type of assay not only facilitates understanding functional genomics of potential targets but also the validation of targets in the human system as described in figure 5.

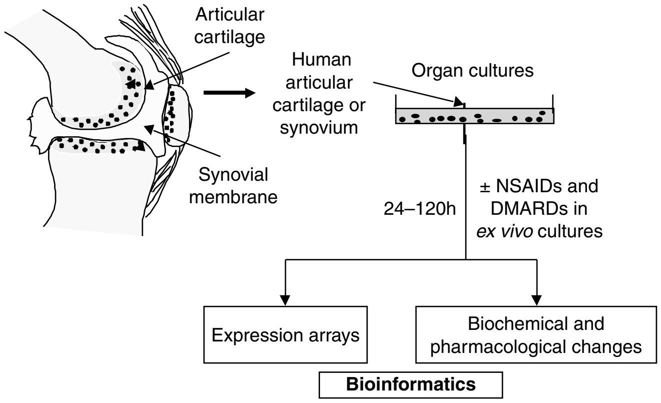

Articular cartilage
Human articular cartilage or synovium
Organ cultures
Synovial membrane
24–120h
± NSAIDs and DMARDs in *ex vivo* cultures
Expression arrays
Biochemical and pharmacological changes
Bioinformatics

Fig. 5. Human arthritis-affected cartilage *ex vivo* assay for functional genomics and pharmacogenomics. Human cartilage organ culture was conducted as described previously.[34] Briefly, knee articular cartilage from patients undergoing knee replacement surgery was obtained and cut into 3mm discs, and 4–6 discs (~100mg) were placed in organ culture in 2mL of Ham's F-12 medium and 0.1% human albumin for 24–72h with or without modulators (recombinant proteins, neutralizing antibodies, small molecule drugs). The medium was analyzed for soluble mediators such as prostaglandin E_2 (PGE_2), nitric oxide, cytokines, and chemokines. RNA was isolated from the cartilage at time zero and different timepoints and utilized for pharmacogenomic analyses such as gene expression arrays. **DMARDs** = disease-modifiying antirheumatic drugs.

Functional Analysis of Osteopontin in Cartilage

Osteopontin (OPN) is a phosphorylated glycoprotein secreted by activated macrophages, T-cell leukocytes and bone derived cells. OPN is present in extracellular fluids at the site of inflammation matrix of mineralized tissues. OPN mediates important cell-cell interactions and cell-matrix interaction through engagement of receptors including integrins $\alpha_5\beta_3$ and CD44 (hyaluronic acid receptor).[95,96] Differential expression of mRNA in normal and arthritis-affected cartilage showed up-regulation (300–700%) of fibronectin and osteopontin.[15,97]

Independently, genetic associations of nodal OA with two loci on chromosome 2q were identified using microsatellite marker loci and sibling pairs. One of the candidate genes in this region was the fibronectin gene (*FN1*), a potential target ECM gene for therapy identified by both differential mRNA expression and conventional population based genetic approaches.[98] Furthermore, Saito et al.[99] have also shown that fragments of fibronectin protein are functionally active as they up-regulate inflammatory mediators in human chondrocytes and other inflammatory cells.

Based on these observations, subsequent studies were focused on the identification of function of osteopontin in OA-affected cartilage. The effect of recombinant osteopontin (or affinity purified anti-osteopontin serum) on chondrocyte function in cartilage was examined. Human recombinant osteopontin at 1 μg/mL (20 nmol/L) inhibited spontaneous or IL-1β-induced NO and PGE_2 production. Both NO and PGE_2 have been reported to be involved in various chondrocytes functions such as apoptosis, MMP activation, growth, and inflammation.[94] Furthermore, neutralization of intra-articular osteopontin with anti-osteopontin serum augmented NO production.[15] These functional studies suggest that one of the roles of intra-articular osteopontin, which is overexpressed in OA cartilage, is to act as an innate inhibitor of IL-1, NO, and PGE_2 production.

Although addition of ECM proteins to cartilage organ cultures causes significant modulation of inflammatory mediators and changes in cartilage homeostasis, one of the main drawbacks of this assay system is that it is not possible to identify the specificity of the matrix ligands, as these ligands have the ability to bind to various receptors on the same cell type.[100,101] For instance, osteopontin can bind to integrins such as $\alpha_v(\beta_1, \beta_3, \text{or } \beta_5)$ dimers and $(\alpha_4, \alpha_5, \alpha_8, \text{or } \alpha_9)\beta_1$ dimers, and also CD44,[102] whereas one of these receptors, $\alpha_v\beta_3$, preferentially binds to osteopontin and to other ligands, such as vitronectin. Thus, it becomes prudent to identify and develop antibodies that competitively inhibit ligand binding (antagonists) and at the same time mimic the function of the ligand (functional agonists) in human cartilage. The $\alpha_5\beta_1$ integrin receptor for fibronectin has been identified on the chondrocytes.[89] Similarly, the $\alpha_v\beta_3$ integrin receptor was also identified on these cell types.[89,103]

The binding of monoclonal antibody JBS5 (which acts as an agonist similar to the N-terminal fragment of fibronectin) to integrin $\alpha_5\beta_1$ up-regulates inflammatory mediators such as NO and PGE_2 as well as the cytokines, IL-6 and IL-8. The $\alpha_v\beta_3$ complex-specific function-blocking monoclonal antibody LM609, which acts as an agonist similar to osteopontin,[102] attenuates the production of IL-1β, NO, and PGE_2 (triggered by $\alpha_5\beta_1$, IL-1β, IL-18, or IL-1β + TNFα + LPS) in a dominant fashion in OA-affected cartilage and in bovine chondrocytes. These data demonstrate a cross-talk in signaling mechanisms among integrins mediated via cartilage matrix components, which also influences 'outside in' and 'inside out' signaling. The dysfunctional signaling mechanisms may influence cartilage homeostasis and its deregulation may play a role in the pathogenesis of OA.[89] This *ex vivo* assay validates ligands and agonist-antagonist interactions in a three-dimensional cartilage system, which may simulate an *in vivo* effect.

© 2004 Adis Data Information BV. All rights reserved.

Identifying the Role of Matrix Proteins by Phenotype-Driven Functional Genomic Studies

Hyperplasia of synoviocytes is observed in patients with RA and OA, despite differences in underlying etiologies of the two disorders. The autosomal recessive disorder CACP affects the joints and shows synoviocyte hyperplasia. Using a positional-candidate gene approach, Marcelino et al.[46] identified mutations in the gene for CACP (*PRG4*), encoding a secreted proteoglycan previously identified as both 'megakaryocyte-stimulating factor precursor' and 'superficial zone protein'. The *PRG4* gene product, proteoglycan 4, contains domains that have homology to somatomedin B, heparin-binding proteins, mucins, and hemopexins.[46] Functional genomic analysis in mice using a *PRG4* knock-out shows similar cartilage destruction and synovial hyperplasia as that observed in patients with CACP with no infiltrating inflammatory cells.[46]

Hurvitz et al.[104] used a positional candidate approach to map the *WISP3* gene, associated with progressive pseudo-rheumatoid dysplasia (PPD). PPD can be initially misdiagnosed as juvenile RA, and, like RA, shows loss of normal cell columnar organization in the growth zone in the subchondral region of cartilage.[91] The *WISP3* gene is a member of the connective tissue growth factor family, which are secreted and matrix-bound. At least nine different *WISP3* mutations were identified in PPD-affected individuals. The normal function of *WISP3* is unknown. Animal knock-out studies may assist in understanding the mechanism of action of *WISP3*.

4.3.3 Cytokines

Tumor necrosis factors (TNF) and IL-1 are the major pro-inflammatory cytokine known to play a role in chronic and destructive arthritis. TNFα is a critical cytokine, which stimulates other pro-inflammatory molecules such as IL-1β, IL-8, monocyte chemotactic protein (MCP) and vascular endothelial growth factor (VEGF). These molecules in turn exacerbate the disease by recruitment of various inflammatory cells into joints, inducing angiogenesis and matrix metallo-proteinases associated with joint disruption. In the last decade, several therapeutic approaches using anti-TNF molecules have been analyzed in RA in both mouse models of RA and human RA patients. Feldmann and Maini[105] have shown in animal models that by systemic administration of anti-TNF molecules such as the anti-TNF antibody infliximab or the soluble TNFα receptor fusion protein etanarcept can reverse the above observed clinical symptoms and signs. Both of these biologics have been approved by the FDA for treating patients with RA and these drugs have shown consistent and remarkable efficiency in some individuals.

Analysis of Interleukin-1 (IL-1) Signaling in Arthritis

The IL-1 family is one of the most potent pro-inflammatory multifunctional cytokines and has a central role in joint inflammation and synovial lining/cartilage destruction. IL-1 activity resides in each of the two molecules IL-1α and IL-1β, which act by binding to a common receptor, IL-1RI, and the signaling component IL-1 receptor accessory protein (IL-1RAcP). A third member of the IL-1 family, the IL-1 receptor antagonist (IL-1ra), also binds IL-1RI but fails to bring about the interaction with IL-1RAcP. Thus, by blocking the receptor, IL-1ra prevents the action of the agonist IL-1. Additional regulation is provided by IL-1 type II (IL-1RII) or decoy receptor, which binds and sequesters IL-1β, thereby preventing signaling.

Binding of IL-1α or IL-1β to IL-1RI forms a heterodimer with IL-1AcP and evokes signal transduction. This leads to activation of the nuclear factor κB (NFκB) pathway. In contrast, IL-1RII, which also binds to IL-1Acp, functions as an antagonist by capturing excess of IL-1 and blocking IL-1 function. Soluble forms of both the receptor types can be found in plasma and can function as molecular sinks for excess circulating IL-1 protein.

Anakinra, a recombinant human form of IL-1ra, saturates IL-1RI and blocks IL-1 signaling. Anakinra has been studied extensively as monotherapy in arthritis.[106] In view of the importance of IL-1 in arthritis,[103] several homologs of IL-1 and its receptors have been identified from the databank using electronic data mining. These include ST2, an IL-1 receptor homolog, and IL-1H1-H4[107-109] as described in figure 6.

Gene array analysis of human normal and arthritis-affected cartilage showed mRNA expression of *IL1RAP* and *IL1R1* (encoding IL-1RAcP and IL-1R1) are down-regulated in OA cartilage, but not *IL1RN* or *IL1R2* (encoding IL-1Ra and IL-1RII, respectively) are not expressed. Similarly, human synovial and epithelial cells also showed low expression of *IL1R2* mRNA[103] (see figure 6).

IL-1 Polymorphisms in Disease Susceptibility and Severity

Differences in disease severity between individuals with OA and RA could be related to different levels of cytokine production or functional differences resulting from polymorphisms in their genes. Genome-wide linkage scans have highlighted up to 7 chromosomal regions that may harbor OA susceptibility genes, of which chromosome 2q is more promising.[112] In a Finnish study of affected sibling pairs, a region of linkage stretching from 2q12 to 2q21 (where the IL-1 receptor and ligand gene clusters reside) was reported for OA.[113] SNP and variable number tandem repeat (VNTR) polymorphisms in the genes *IL1A*, *IL1B* (encoding IL-1α and IL-1β, respectively) and *IL1RN* have been associated with chronic inflammatory diseases, autoimmune diseases, and gastric cancer.[114,115] Recently, Loughlin et al.[116] genotyped 7 SNPs and 1 VNTR from within the genes of the IL-1 cluster and showed a marginal association of these with OA. Independently, Moos et al.[117] also reported an association between the IL-1 ligand gene cluster and OA. Amin, Thakker and colleagues[118] have reported promoter analysis and IL-1β expression in human chondrocytes,

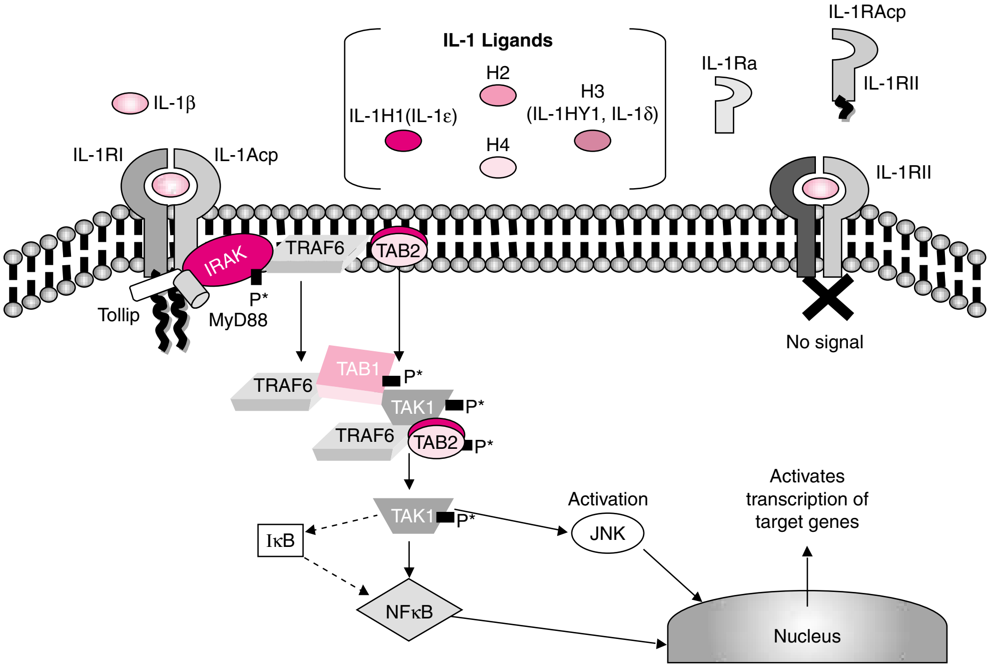


Fig. 6. Interleukin (IL)-1 receptors and signaling mechanisms. IL-1 signaling is regulated by the receptors IL-1 receptor type I (IL-1RI) and IL-1RII. The effects of IL-1 receptors are mediated by IL-1α and IL-1β. IL-1 receptor antagonist (IL-1Ra, anakinra) blocks signaling through the receptor. Accessory proteins (Acp) and receptors interact with IL-1RI during the initial phase of IL-1R-ligand interaction and signal transduction.[110] IL-1RII can also inhibit IL-1 signaling in a similar fashion. Binding of ligands to IL-1RI leads to formation of a complex comprising IL-1RI, IL-1 receptor accessory protein (IL-1AcP), Toll interacting protein (Tollip) and myeloid differentiation factor 88 (MyD88). This complex recruits IL-1 receptor associated kinases (IRAKs) to the inner membrane where it gets phosphorylated, and forms a complex with tumor necrosis factor receptor (TNFR)-associated factor 6 (TRAF6). TRAF6 then translocates from the membrane to cytosol. Membrane bound IRAK also facilitates translocation of IL-1-induced transforming growth factor β (TNFβ) activated kinase 1 binding protein 2 (TAB2) from membrane to cytosol. Translocation of TAB2 and TRAF6 leads to formation of a TRAF6-TAK1-TAB1-TAB2 complex in the cytosol resulting in the activation of TGFβ-activated kinase 1 (TAK1) and, eventually, in IL-1-induced NFκB activation. IL-1β has also been reported to induce p38 MAP kinase, c-Jun-N terminal kinase and p42/p44 ERK MAP kinase in chondrocytes.[111] In contrast, cell-surface IL-1RII (when engaged with IL-1α or IL-1β) binds to soluble IL-1RAcp (IL-1sAcp) and sequesters it, competitively inhibiting the accessory protein from binding to *signaling* IL-1RI. This decoy receptor complex (unlike IL-1RI) does not mediate a signal. The soluble form of IL-1 type II decoy receptor (IL-1RII) is active and binds to IL-1. **IL-1H1-H4** = IL-1 homologos 1-4; **IκB** = inhibitor of NFκB; **TAB1** and **TAB2** = transforming growth factor B activating kinase binding protein.

suggesting the central role of the upstream IL-1 enhancer in regulation of IL-1 genes by IL-1.

In summary, there are significant independent genomic-based reports that suggest that IL-1β may be one of the important players contributing to the effects observed in the joints.

Gene Transfer Approach for Functional Genomics *in Vitro*

The above observations suggest that low concentrations of IL-1β (picogram amounts of IL-1 in gram amount of cartilage), released in human OA-affected cartilage during early stages of the disease,[103] has the ability to act unopposed with respect to the lack of naturally occurring IL-1 antagonistic activity in the cartilage. Thus, low concentrations of functional IL-1 can have detrimental effects on cartilage homeostasis in long-term diseases such as OA. In view of our previous observations,[103] the function of the membrane and soluble form of IL-1RII in human chondrocyte and cartilage homeostasis was examined by reconstituting the receptor in IL-1RII-deficient chondrocytes, synovial cells and epithelial fibroblasts across various species.[103]

Functional genomic analysis showed that recombinant soluble IL-1RII (sIL-1RII), at picomolar concentrations, but not soluble TNF receptor-IgG Fc fusion protein, significantly inhibited IL-1β-induced NO and/or PGE$_2$ production in chondrocytes, synovial and epithelial cells. In OA-affected cartilage, the dose which produces 50% inhibition (IC$_{50}$) of NO production by sIL-1RII was one log order lower than that for sIL-1RI.[103]

Reconstitution of human IL-1RII expression in various IL-1RII-deficient cell types was examined to determine its regulation, specificity and antagonist role using gene transfer approaches. Adenovirus expressing human IL-1RII (AdRSVIL-1RII), when transduced into human and bovine chondrocytes, synovial cells and epithelial cells expressed membrane IL-1RII (mIL-1RII) and spontaneously released functional soluble IL-1RII (sIL-1RII) in the culture supernatants. The IL-1RII$^+$ (but not IL-1RII$^-$) cells were resistant to autocrine and exogenous IL-1-induced effects of

© 2004 Adis Data Information BV. All rights reserved.

NO, PGE_2, IL-6, and IL-8, or decrease in proteoglycan synthesis. Human IL-1RII was functional in various cell types and across species. IL-1RII is cell specific in chondrocytes but also regulates TNFα induced inflammatory mediators in human synovial and epithelial cells. $IL\text{-}1RII^+$ chondrocytes were more resistant to induction of NO and PGE_2 by IL-1 as compared with $IL\text{-}1RII^-$ cells incubated with 100-fold excess of sIL-1RII protein. In cocultures, $IL\text{-}1RII^+$ synovial cells released a functional sIL-1RII, which in a paracrine fashion protected chondrocytes from the effects of IL-1. Furthermore, autologous $IL\text{-}1RII^+$ (but not $IL\text{-}1RII^-$) chondrocytes, when transplanted onto human OA-cartilage *in vitro* showed spontaneous release of sIL-1RII for 20 days, and inhibited the spontaneous production of NO and PGE_2 in cartilage in *ex vivo* conditions.

In summary, reconstitution of IL-1RII in $IL\text{-}1RII^-$ cells using viral and/or plasmid expression vectors significantly protects $IL\text{-}1RII^-$ cells against the autocrine/paracrine effects of IL-1β by acting at several levels of IL-1 signaling and transcription.[103]

Other Cytokines: An Adenoviral-Based Approach to Functional Genomics *in Vivo*

Polymorphism in the *IL1B* gene is associated with inflammation in arthritis,[119] TGFβ1 gene (*TGFB1*) is associated with spinal OA (osteophytosis) and the insulin-like growth factor (IGF) gene (*IGF1*) is associated with generalized OA.[119] Differential expression of mRNA in normal and arthritis-affected cartilage also showed upregulation of *IL1*β, *TGFB1*, and *IGF1* mRNA transcripts.[119,120] The role of IL-1, TGFβ, and IGF in joints can be assessed using a gene therapy approach in a collagen-induced mouse arthritis model and rabbit models to dissect the functions of these growth factors in joint destruction and cartilage homeostasis.

Constitutive intra-articular expression of an adenoviral *IL1B* transgene in rabbit joints induces multiple intra-articular manifestations, which include intense inflammation, leukocytosis, synovial hypertrophy, hyperplasia, highly aggressive pannus formation, erosion of cartilage and bone. It also induced systemic effects including diarrhea and fever. Following the loss of the transgene, (which occurs after 28 days) most of the pathophysiological symptoms described above recur within 4 weeks.[121]

Over-expression of TGFβ by adenovirus in rodent joints induced the formation of osteophytes [122] and deregulation of bone remodeling,[123] in spite of some of its beneficial effects on chondrocyte metabolism.[123] Gene transfer of *IGF1* into rabbit knee joints promotes proteoglycan synthesis without significantly affecting inflammation or cartilage breakdown. No adverse effect of IGF expression was observed on synovium or adjacent muscles. This local gene transfer of *IGF1* to joints could serve as a therapeutic strategy to stimulate new matrix synthesis in both RA and OA.[124] Other target genes (*TSG*-6, *IL4*, *SOD*, *IL1RN*, *IL1RII*, and $p16^{INK4a}$) have also been tested in this model system with success.[125-128]

In summary, a gene therapy approach (*in vitro* or *in vivo*) allows identification of the function of candidate genes in the complex cartilage environment.

5. Conclusions and Future Directions

A simple association of gene expression with disease (e.g. generated through a gene array) does not validate a role for that target gene in the disease. Nor does a human genetic approach to identify targets associated with the disease necessarily generate chemically-tractable molecular targets. Rather the goal of target validation and functional genomics is to strengthen correlative data (from gene arrays, EST libraries, and proteomics) by demonstrating a definitive role for the candidate gene in a disease model. Bioinformatics is an important tool in functional genomics that will help bridge the gap between correlative data and causative data, although there are limitations in predicting *ab initio* gene structure, gene function and protein folds from the raw sequence data. Clearly, much needs to be done, as more than 40% of the 35 000 genes (and possibly 120 000 different proteins they may code) have not been ascribed any functional attribute,[129] either a biochemical function (e.g. kinase), cellular function (e.g. a specific signaling pathway), or function at the organism level (e.g. brain development, immune response, etc.).

As of 1996, successful drug treatments involved fewer than 500 targets or growth factors.[130] It is assumed that at least 5000 of the possible 120 000 proteins may be potential therapeutic proteins or targets, suggesting that only 10% of potential therapeutic strategies have been identified and exploited to date.[131] The gene(s) involved in the etiology of arthritis, subcategorized based on their clinical symptoms, still remain to be identified and characterized among these targets. Genome-based translational research efforts coupled with bioinformatics will facilitate correlating gene expression data with clinical outcomes as described in normal and RA-affected monocytes.[132] These preliminary studies may lead to predictive medicine in arthritis, identifying different disease states, and biomarkers (e.g. therapy-induced remission), with respect to modulation of novel genes.

Arthritis is currently managed with a variety of drugs. On the whole, these drugs treat inflammation and pain as symptoms, but do not address the actual cause of the disease. Some of the new generation of therapeutics in development include vascular adhesion protein 1 (VAP-1) and vascular endothelial growth factor (VEGF) and its receptor FIt-1, which aim to treat the symptoms of the disease. Gene therapy approaches in human RA have given some promising results in phase II clinical trials.[133] Increasing the understanding of molecular cascades involved in the disease processes by genomic approaches will allow us to produce significantly better drugs and biomarkers for predictive medicine than in the future with increased selectivity and fewer adverse effects.

© 2004 Adis Data Information BV. All rights reserved.

Acknowledgements

We thank Ms Andrea L. Barrett and Ms Cari Reiner for the preparation of the manuscript and Ms Sonali Trivedi for preparation of the figures. We would like to thank Dr Smita Palejwala (University of Medicine and Dentistry of New Jersey) for critically reviewing the manuscript and for her constructive suggestions. We would like to thank NDRI (Philadelphia, PA) for some of the cartilage samples.

We would like to thank the American Association of Orthopedic Surgeons and the Journal of Immunology for allowing us to republish figures 1 and 4, respectively.

References

1. McCarty DJ. Differential diagnosis of arthritis: analysis of signs and symptoms. In: Koopman WJ, editor. Arthritis and allied conditions: a textbook of rheumatology. Philadelphia (PA): Lippincott, 1998: 39-50
2. Riley JH, Allan CJ, Lai E, et al. The use of single nucleotide polymorphisms in the isolation of common disease genes. Pharmacogenomics 2000; 1: 39-47
3. Jimenez-Sanchez G, Childs B, Valle D. Human disease genes. Nature 2001; 409: 853-5
4. Peltonen L, McKusick VA. Genomics and medicine: dissecting human disease in the postgenomic era. Science 2001; 291: 1224-9
5. Frary A, Nesbitt TC, Grandillo S, et al. Fw2.2: a quantitative trait locus key to the evolution of tomato fruit size. Science 2000; 289: 85-8
6. Masinde GL, Li X, Gu W, et al. Identification of wound healing/regeneration quantitative trait loci (QTL) at multiple time points that explain seventy percent of variance in (MRL/MpJ and SJL/J) mice F2 population. Genome Res 2001; 11: 2027-33
7. Brockmann GA, Bevova MR. Using mouse models to dissect the genetics of obesity. Trends Genet 2002; 18: 367-76
8. Korstanje R, Paigen B. From QTL to gene: the harvest begins. Nat Genet 2002; 31: 235-6
9. Hamet P, Tremblay J. Genetic determinants of the stress response in cardiovascular disease. Metabolism 2002; 51: 15-24
10. Griffiths MM, Wang J, Joe B, et al. Identification of four new quantitative trait loci regulating arthritis severity and one new quantitative trait locus regulating autoantibody production in rats with collagen-induced arthritis. Arthritis Rheum 2000; 43: 1278-89
11. Joe B, Remmers EF, Dobbins DE, et al. Genetic dissection of collagen-induced arthritis in Chromosome 10 quantitative trait locus speed congenic rats: evidence for more than one regulatory locus and sex influences. Immunogenetics 2000; 51: 930-44
12. Gulko PS, Kawahito Y, Remmers EF, et al. Identification of a new non-major histocompatibility complex genetic locus on chromosome 2 that controls disease severity in collagen-induced arthritis in rats. Arthritis Rheum 1998; 41: 2122-31
13. Kawahito Y, Cannon GW, Gulko PS, et al. Localization of quantitative trait loci regulating adjuvant-induced arthritis in rats: evidence for genetic factors common to multiple autoimmune diseases. J Immunol 1998; 161: 4411-9
14. Patel IR, Attur MG, Patel RN, et al. TNF-α convertase from human arthritis-affected cartilage: isolation of CDNA by differential display, expression of the active enzyme, and regulation of TNF-α. J Immunol 1998; 160: 4570-9
15. Attur MG, Dave MN, Stuchin S, et al. Osteopontin: an intrinsic inhibitor of inflammation in cartilage. Arthritis Rheum 2001; 44: 578-84
16. Hu SI, Carozza M, Klein M, et al. Human HtrA, an evolutionarily conserved serine protease identified as a differentially expressed gene product in osteoarthritic cartilage. J Biol Chem 1998; 273: 34406-12
17. Heller RA, Schena M, Chai A, et al. Discovery and analysis of inflammatory disease-related genes using CDNA microarrays. Proc Natl Acad Sci U S A 1997; 94: 2150-5
18. Khan IM, Salter DM, Bayliss MT, et al. Expression of clusterin in the superficial zone of bovine articular cartilage. Arthritis Rheum 2001; 44: 1795-9
19. Kumar S, Connor JR, Dodds RA, et al. Identification and initial characterization of 5000 expressed sequenced tags (ESTs) each from adult human normal and osteoarthritic cartilage CDNA libraries. Osteoarthritis Cartilage 2001; 9: 641-53
20. Justen HP, Grunewald E, Totzke G, et al. Differential gene expression in synovium of rheumatoid arthritis and osteoarthritis. Mol Cell Biol Res Commun 2000; 3: 165-72
21. Seki T, Selby J, Haupl T, et al. Use of differential subtraction method to identify genes that characterize the phenotype of cultured rheumatoid arthritis synoviocytes. Arthritis Rheum 1998; 41: 1356
22. Amin AR. Dissecting human musculoskeletal diseases in the post-genomic era of medicne: following the scent in search for the the etiological gene(s). In: Rosier RN, Evans CH, editor. Molecular biology in orthopaedic. Rosemont (IL): American Academy of Orthopaedic Surgeons, 2003: 3-12
23. Attur MG, Dave MN, Tsunoyama K, et al. A "system biology" approach to bioinformatics and functional genomics in complex human diseases: arthritis. Curr Issues Mol Biol 2002; 4: 129-46
24. Attur MG, Bingham III CO, Dave MN, et al. Model protocol to study pharmacogenomics in inflammatory diseases: human rheumatoid arthritis. Drug Dev Res 2000; 49: 29-33
25. Rioux JD, Daly MJ, Silverberg MS, et al. Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease. Nat Genet 2001; 29: 223-8
26. Ingvarsson T, Stefansson SE, Gulcher JR, et al. A large icelandic family with early osteoarthritis of the hip associated with a susceptibility locus on chromosome 16p. Arthritis Rheum 2001; 44: 2548-55
27. Jawaheer D, Seldin MF, Amos CI, et al. A genomewide screen in multiplex rheumatoid arthritis families suggests genetic overlap with other autoimmune diseases. Am J Hum Genet 2001; 68: 927-36
28. Knight J. When the chips are down. Nature 2001; 410: 860-1
29. Hastie T, Tibshirani R, Eisen MB, et al. 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. Genome Biol 2000; 1: RESEARCH0003. Epub 2000 aug 04.
30. Bassett Jr DE, Eisen MB, Boguski MS. Gene expression informatics: it's all in your mine. Nat Genet 1999; 21 (1 Suppl.): 51-5
31. Rivera MC, Jain R, Moore JE, et al. Genomic evidence for two functionally distinct gene classes. Proc Natl Acad Sci U S A 1998; 95: 6239-44
32. Eisen MB, Spellman PT, Brown PO, et al. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 1998; 95: 14863-8
33. Attur MG, Dave M, Akamatsu M, et al. Osteoarthritis or osteoarthrosis: the definition of inflammation becomes a semantic issue in the genomic era of molecular medicine. Osteoarthritis Cartilage 2002; 10: 1-4
34. Attur MG, Dave MN, Leung MY, et al. Functional genomic analysis of type II IL-1β decoy receptor: potential for gene therapy in human arthritis and inflammation. J Immunol 2002; 168: 2001-10
35. Wang FL, Connor JR, Dodds RA, et al. Differential expression of Egr-1 in osteoarthritic compared to normal adult human articular cartilage. Osteoarthritis Cartilage 2000; 8: 161-9
36. Tetlow LC, Adlam DJ, Woolley DE. Matrix metalloproteinase and proinflammatory cytokine production by chondrocytes of human osteoarthritic cartilage: associations with degenerative changes. Arthritis Rheum 2001; 44: 585-94
37. Freemont AJ, Hampson V, Tilman R, et al. Gene expression of matrix metalloproteinases 1, 3, and 9 by chondrocytes in osteoarthritic human knee articular cartilage is zone and grade specific. Ann Rheum Dis 1997; 56: 542-9
38. Goldring MB, Birkhead JR, Suen L-F, et al. Interleukin-1β–modulated gene expression in immortalized human chondrocytes. J Clin Invest 1994; 94: 2307-16
39. Hauselmann HJ, Aydelotte MB, Schumacher BL, et al. Synthesis and turnover of proteoglycans by human and bovine adult articular chondrocytes cultured in alginate beads. Matrix 1992; 12: 116-29
40. Lemare F, Steimberg N, Le Griel C, et al. Dedifferentiated chondrocytes cultured in alginate beads: restoration of the differentiated phenotype and of the metabolic responses to interleukin-1beta. J Cell Physiol 1998; 176: 303-13
41. Perka C, Spitzer RS, Lindenhayn K, et al. Matrix-mixed culture: new methodology for chondrocyte culture and preparation of cartilage transplants. J Biomed Mater Res 2000; 49: 305-11
42. Yang BB, Zhang Y, Cao L, et al. Aggrecan and link protein affect cell adhesion to culture plates and to type II collagen. Matrix Biol 1998; 16: 541-61
43. Aigner T, Dertinger S, Vornehm SI, et al. Phenotypic diversity of neoplastic chondrocytes and extracellular matrix gene expression in cartilaginous neoplasms. Am J Pathol 1997; 150: 2133-41
44. Haas C, Aicher WK, Dinkel A, et al. Characterization of SV40T antigen immortalized human synovial fibroblasts: maintained expression patterns of EGR-1, HLA-DR and some surface receptors. Rheumatol Int 1997; 16: 241-7

© 2004 Adis Data Information BV. All rights reserved.

45. Rosier RN, Evans CH, editor. Molecular biology in orthopaedic. Rosemont (IL): American Academy of Orthopaedic Surgeons, 2002: 107-14
46. Marcelino J, Carpten JD, Suwairi WM, et al. CACP, encoding a secreted proteoglycan, is mutated in camptodactyly-arthropathy-coxa vara-pericarditis syndrome. Nat Genet 1999; 23: 319-22
47. Chan D, Weng YM, Graham HK, et al. A nonsense mutation in the carbosyl-terminal domain of type X collagen causes haploinsufficiency in schmid metaphyseal chondrodysplasia. J Clin Invest 1998; 101: 1490-9
48. Martignetti JA, Aqeel AA, Sewairi WA, et al. Mutation of the matrix metalloproteinase 2 gene (MMP2) causes a multicentric osteolysis and arthritis syndrome. Nat Genet 2001; 28: 261-5
49. Matsumoto I, Staub A, Benoist C, et al. Arthritis provoked by linked T and B cell recognition of a glycolytic enzyme. Science 1999; 286: 1732-5
50. Jorgensen C, Gay S. Gene therapy in osteoarticular diseases: where are we? Immunol Today 1998; 19: 387-91
51. Muller-Ladner U, Roberts CR, Franklin BN, et al. Human IL-1Ra gene transfer into human synovial fibroblasts is chondroprotective. J Immunol 1997; 158: 3492-8
52. Muller-Ladner U, Evans CH, Franklin BN, et al. Gene transfer of cytokine inhibitors into human synovial fibroblasts in the SCID mouse model. Arthritis Rheum 1999; 42: 490-7
53. Arita M, Li SW, Kopen G, et al. Skeletal abnormalities and ultrastructural changes of cartilage in transgenic mice expressing a collagen II gene (COL2A1) with a Cys for Arg-alpha1-519 substitution. Osteoarthritis Cartilage 2002; 10: 808-15
54. Gustafsson E, Aszodi A, Ortega N, et al. Role of collagen type II and perclean in skeletal development. Ann N Y Acad Sci 2003; 995: 140-50
55. Helminen HJ, Saamanen AM, Salminen H, et al. Transgenic mouse models for studying the role of cartilage macromolecules in osteoarthrists. Rheumatology (Oxf) 2002; 41: 848-56
56. Mier RJ, Holderbaum D, Ferguson R, et al. Osteoarthritis in children associated with a muation in the type II procollagen gene (COL2A1). Mol Genet Metab 2001; 74; 338-41
57. Baldwin CT, Cupples LA, Joost O, et al. Absence of linkage or association for osteoarthritis with the vitamin D receptor/type II collagen locus: the Framingham Osteoarthritis Study. J Rheumatol 2002; 29: 161-5
58. Petit B, Freyria AM, van der Rest M, et al. Cartilage collagens. In: Adolphe M, editor. Biological regulation of the chondrocytes. Boca Raton (FL): CRC Press, 1998: 33-72
59. Kuivaniemi H, Tromp G, Prockop DJ. Mutations in fibrillar collagens (Types I, II, III, and XI), fibril-associated collagen (Type IX), and network-forming collagen (type X) cause a spectrum of diseases of bone, cartilage, and blood vessels. Hum Mutat 1997; 9: 300-15
60. Fassler R, Schnegelsberg PN, Dausman J, et al. Mice lacking alpha 1 (IX) collagen develop noninflammatory degenerative joint disease. Proc Natl Acad Sci U S A 1994; 91: 5070-4
61. Nakata K, Ono K, Miyazaki J, et al. Osteoarthritis associated with mild chondrodysplasia in transgenic mice expressing alpha 1(IX) collagen chains with a central deletion. Proc Natl Acad Sci U S A 1993; 90: 2870-4
62. Aszodi A, Hunziker EB, Olsen BR, et al. The role of collagen II and cartilage fibril-associated molecules in skeletal development. Osteoarthritis Cartilage 2001; 9 Suppl. A: S150-9
63. Childs S, Chen JN, Garrity DM, et al. Patterning of angiogenesis in the zebrafish embryo. Development 2002; 129: 973-82
64. Westermarck J, Kahari VM. Regulation of matrix metalloproteinase expression in tumor invasion. FASEB J 1999; 13: 781-92
65. Chandler S, Coates R, Gearing A, et al. Matrix metalloproteinases degrade myelin basic protein. Neurosci Lett 1995; 201: 223-6
66. Rosenberg GA. Matrix metalloproteinases in brain injury. J Neurotrauma 1995; 12: 833-42
67. George SJ. Therapeutic potential of matrix metalloproteinase inhibitors in atherosclerosis. Expert Opin Investig Drugs 2000; 9: 993-1007
68. Longo GM, Xiong W, Greiner TC, et al. Matrix metalloproteinases 2 and 9 work in concert to produce aortic aneurysms. J Clin Invest 2002; 110: 625-32
69. Shahin M, Konturek JW, Pohle T, et al. Remodeling of extracellular matrix in gastric ulceration. Microsc Res Tech 2001; 53: 396-408
70. Boeker KH, Haberkorn CI, Michels D, et al. Diagnostic potential of circulating TIMP-1 and MMP-2 as markers of liver fibrosis in patients with chronic hepatitis C. Clin Chim Acta 2002; 316: 71-81
71. Mengshol JA, Mix KS, Brinckerhoff CE. Matrix metalloproteinases as therapeutic targets in arthritic diseases: bull's-eye or missing the mark? Arthritis Rheum 2002; 46: 13-20
72. Bayliss MT, Hutton S, Hayward J, et al. Distribution of aggrecanase (ADAMts 4/5) cleavage products in normal and osteoarthritic human articular cartilage: the influence of age, topography and zone of tissue. Osteoarthritis Cartilage 2001; 9: 553-60
73. Brinckerhoff CE, Matrisian LM. Matrix metalloproteinases: a tail of a frog that became a prince. Nat Rev Mol Cell Biol 2002; 3: 207-14
74. Martel-Pelletier J, Welsch DJ, Pelletier JP. Metalloproteases and inhibitors in arthritic diseases. Best Pract Res Clin Rheumatol 2001; 15: 805-29
75. Cawston T, Billington C, Cleaver C, et al. The regulation of MMPs and TIMPs in cartilage turnover. In: Greenwald RA, Zucker S, Golub LM, editors. Inhibition of matrix metalloproteinases therapeutic applications. New York: The New York Academy of Sciences, 1999: 120-9
76. Sternlicht MD, Werb Z. How matrix metalloproteinasesregulate cell behavior. Annu Rev Cell Dev Biol 2001; 17: 463-516
77. Lohi J, Wilson CL, Roby JD, et al. Epilysin, a novel human matrix metalloproteinase (MMP-28) expressed in testis and keratinocytes and in response to injury. J Biol Chem 2001; 276: 10134-44
78. Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. Nature 2001; 409: 860-921
79. Venter JC, Adams MD, Myers EW, et al. The sequence of the human genome. Science 2001; 291: 1304-51
80. Wolfsberg TG, Bazan JF, Blobel CP, et al. The precursor region of a protein active in sperm-egg fusion contains a metalloprotease and a disintegrin domain: structural, functional, and evolutionary implications. Proc Natl Acad Sci U S A 1993; 90: 10783-7
81. Abagyan RA, Totrov MM, Kuznetsov DA. ICM, a new method for structure modeling and design: applications to docking and structure prediction from the distorted native conformation. J Comp Chem 1994; 15: 488-506
82. Moss ML, White JM, Lambert MH, et al. TACE and other ADAM proteases as targets for drug discovery. Drug Discov Today 2001; 6: 417-26
83. Attur MG, Dave MN, Patel IR, et al. Regulation of inflammatory mediators by tetracyclines. In: Nelson M, Hillen W, Greenwald RA, editors. Tetracyclines in biology, chemistry and medicine. Boston (MA): Birkhauser Verlag, 2001: 295-310
84. Black RA, Rauch CT, Kozlosky CJ, et al. A metalloproteinase disintegrin that releases tumor-necrosis factor-α from cells. Nature 1997; 385: 729-33
85. Ganu V, Melton, R, Hu S, et al. Extracellular matrix proteins COMP and fibronectin are substrates for Htra, a novel serine protease upregulated in osteoarthritic cartilage. 47th Annual Meeting, Orthopaedic Research Society; 2001 Feb 25-28; San Francisco.
86. Holmbeck K, Bianco P, Caterina J, et al. MT1-MMP-deficient mice develop dwarfism, osteopenia, arthritis, and connective tissue disease due to inadequate collagen turnover. Cell 1999; 99: 81-92
87. Bakker AC. van de Loo, FA, van Beuningen HM, et al. Overexpression of active TGF-beta-1 in the murine knee joint: evidence for synovial-layer-dependent chondro-osteophyte formation. Osteoarthritis Cartilage 2001; 9: 128-36
88. Loeser RF. Chondrocyte integrin expression and function. Biorheology 2000; 37: 109-16
89. Attur MG, Dave MN, Clancy RM, et al. Functional genomic analysis in arthritis-affected cartilage: Yin-Yang regulation of inflammatory mediators by alpha5beta1 and alphaVbeta3 integrins. J Immunol 2000; 164: 2684-91
90. Heinegard D, Bayliss M, Lorenzo P. Pathogenesis of osteoarthritis. In: Brandt KD, Doherty M, Lohmander LS, editors. Osteoarthritis. Oxford: Oxford University Press, 1998: 74-84
91. Attur MG, Patel IR, Patel RN, et al. Autocrine production of IL-1β by human osteoarthritis-affected cartilage and differential regulation of endogenous nitric oxide, IL-6, prostaglandin E_2 and IL-8. Proc Assoc Am Physicians 1998; 110: 1-8
92. Amin AR, Attur MG, Abramson SB. Regulation of nitric oxide and inflammatory mediators in human osteoarthritis-affected cartilage: implication for pharmacological intervention. In: Rubanyi GM, editor. The pathophysiology and clinical applications of nitric oxide. Richmond (CA): Harwood Academic Publishers, 1999: 397-413
93. Amin AR, Attur M, Abramson SB. Nitric oxide synthase and cyclooxygenases: distribution, regulation, and intervention in arthritis. Curr Opin Rheumatol 1999; 11: 202-9

© 2004 Adis Data Information BV. All rights reserved.

94. Amin AR, Dave M, Attur M, et al. Cox-2, no, and cartilage damage and repair. Curr Rheumatol Rep 2000; 2: 447-53

95. Sodek J, Ganss B, McKee MD. Osteopontin. Crit Rev Oral Biol Med 2000; 11: 279-303

96. Sodek J, Zhu B, Huynh MH, et al. Novel functions of the matricellular proteins osteopontin and osteonectin/SPARC. Connect Tissue Res 2002; 43: 308-19

97. Pullig O, Weseloh G, Gauer S, et al. Osteopontin is expressed by adult human osteoarthritic chondrocytes: protein and MRNA analysis of normal and osteoarthritic cartilage. Matrix Biol 2000; 19: 245-55

98. Wright GD, Hughes AE, Regan M, et al. Association of two loci on chromosome 2q with nodal osteoarthritis. Ann Rheum Dis 1996; 55: 317-9

99. Saito S, Yamaji N, Yasunaga K, et al. The fibronectin extra domain a activates matrix metalloproteinase gene expression by an interleukin-1-dependent mechanism. J Biol Chem 1999; 274: 30756-63

100. Aplin AE, Howe A, Alahari SK, et al. Signal transduction and signal modulation by cell adhesion receptors: the role of integrins, cadherins, immunoglobulin-cell adhesion molecules, and selectins. Pharmacol Rev 1998; 50: 197-263

101. Longhurst CM, Jennings LK. Integrin-mediated signal transduction. Cell Mol Life Sci 1998; 54: 514-26

102. Denhardt DT, Noda M, O'Regan AW, et al. Osteopontin as a means to cope with environmental insults: regulation of inflammation, tissue remodeling, and cell survival. J Clin Invest 2001; 107: 1055-61

103. Attur MG, Dave M, Cipolletta C, et al. Reversal of autocrine and paracrine effects of interleukin 1 (IL-1) in human arthritis by type II IL-1 decoy receptor: potential for pharmacological intervention. J Biol Chem 2000; 275: 40307-15

104. Hurvitz JR, Suwairi WM, Van Hul W, et al. Mutations in the CCN gene family member WISP3 cause progressive pseudorheumatoid dysplasia. Nat Genet 1999; 23: 94-8

105. Feldmann M, Maini RN. Links discovery of TNF-alpha as a therapeutic target in rheumatoid arthritis: preclinical and clinical studies. Joint Bone Spine 2002 Jan; 69 (1): 12-8

106. Cohen S, Hurd E, Cush J, et al. Treatment of rheumatoid arthritis with anakinra, a recombinant human interleukin-1 receptor antagonist, in combination with methotrexate: results of a twenty-four-week, multicenter, randomized, double-blind, placebo-controlled trial. Arthritis Rheum 2002; 46: 614-24

107. Lin H, Ho AS, Haley-Vicente D, et al. Cloning and characterization of IL-1Hy2, a novel interleukin-1 family member. J Biol Chem 2001; 276: 20597-602

108. Kumar S, McDonnell PC, Lehr R, et al. Identification and initial characterization of four novel members of the interleukin-1 family. J Biol Chem 2000; 275: 10308-14

109. Mulero JJ, Pace AM, Nelken ST, et al. IL1HY1: a novel interleukin-1 receptor antagonist gene. Biochem Biophys Res Commun 1999; 263: 702-6

110. O'Neill LA, Greene C. Signal transduction pathways activated by the IL-1 receptor family: ancient signaling machinery in mammals, insects, and plants. J Leukoc Biol 1998; 63: 650-7

111. Geng Y, Valbracht J, Lotz M. Selective activation of the mitogen-activated protein kinase subgroups c-Jun NH2 terminal kinase and p38 by IL-1 and TNF in human articular chondrocytes. J Clin Invest 1996; 98: 2425-30

112. Loughlin J. Genetic epidemiology of primary osteoarthritis. Curr Opin Rheumatol 2001; 13: 111-6

113. Leppavuori J, Kujala U, Kinnunen J, et al. Genome scan for predisposing loci for distal interphalangeal joint osteoarthritis: evidence for a locus on 2q. Am J Hum Genet 1999; 65: 1060-7

114. Haukim N, Bidwell JL, Smith AJ, et al. Cytokine gene polymorphism in human disease: on-line databases, Supplement 2. Genes Immun 2002; 3: 313-30

115. El Omar EM, Carrington M, Chow WH, et al. Interleukin-1 polymorphisms associated with increased risk of gastric cancer. Nature 2000; 404: 398-402

116. Loughlin J, Dowling B, Mustafa Z, et al. Association of the interleukin-1 gene cluster on chromosome 2q13 with knee osteoarthritis. Arthritis Rheum 2002; 46: 1519-27

117. Moos V, Rudwaleit M, Herzog V, et al. Association of genotypes affecting the expression of interleukin-1beta or interleukin-1 receptor antagonist with osteoarthritis [published erratum in Arthritis Rheum 2001; 44 (7): 1715]. Arthritis Rheum 2000; 43: 2417-22

118. Amin AR, Thakker G, Patel IR, et al. Regulation of IL-1 gene transcription in human chondrocytes and macrophages: preferential susceptibility of chondrocytes to IL-1 and diverse signaling pathways [abstract]. Arthritis Rheum 2000; 43 Suppl. S: 599

119. Meulenbelt I, Bijkerk C, Miedema HS, et al. A genetic association study of the IGF-1 gene and radiological osteoarthritis in a population-based cohort study (the Rotterdam Study). Ann Rheum Dis 1998; 57: 371-4

120. Yamada Y, Okuizumi H, Miyauchi A, et al. Association of transforming growth factor beta1 genotype with spinal osteophytosis in japanese women. Arthritis Rheum 2000; 43: 452-60

121. Ghivizzani SC, Kang R, Georgescu HI, et al. Constitutive intra-articular expression of human il-1 beta following gene transfer to rabbit synovium produces all major pathologies of human rheumatoid arthritis. J Immunol 1997; 159: 3604-12

122. Van den Berg WB. Growth factors in experimental osteoarthritis: transforming growth factor beta pathogenic? J Rheumatol Suppl 1995; 43: 143-5

123. Smith P, Shuler FD, Georgescu HI, et al. Genetic enhancement of Matrix synthesis by articular chondrocytes: comparison of different growth factor genes in the presence and absence of interleukin-1. Arthritis Rheum 2000; 43: 1156-64

124. Mi Z, Ghivizzani SC, Lechman ER, et al. Adenovirus-mediated gene transfer of insulin-like growth factor 1 stimulates proteoglycan synthesis in rabbit joints. Arthritis Rheum 2000; 430: 2563-70

125. Woods JM, Katschke KJ, Volin MV, et al. IL-4 adenoviral gene therapy reduces inflammation, proinflammatory cytokines, vascularization, and bony destruction in rat adjuvant-induced arthritis. J Immunol 2001; 166: 1214-22

126. Bardos T, Kamath RV, Mikecz K, et al. Anti-inflammatory and chondroprotective effect of TSG-6 (tumor necrosis factor-alpha-stimulated gene-6) in murine models of experimental arthritis. Am J Pathol 2001; 159: 1711-21

127. Taniguchi K, Kohsaka H, Inoue N, et al. Induction of the P16INK4a Senescence gene as a new therapeutic strategy for the treatment of rheumatoid arthritis. Nat Med 1999; 5: 760-7

128. Iyama S, Okamoto T, Sato T, et al. Treatment of murine collagen-induced arthritis by ex vivo extracellular superoxide dismutase gene transfer. Arthritis Rheum 2001; 44: 2160-7

129. Yaspo ML. Taking a functional genomics approach in molecular medicine. Trends Mol Med 2001; 7: 494-501

130. Bumol TF, Watanabe AM. Genetic information, genomic technologies, and the future of drug discovery. JAMA 2001; 285: 551-5

131. Drews J. Drug discovery: a historical perspective. Science 2000; 287: 1960-4

132. Stuhlmuller B, Ungethum U, Scholze S, et al. Identification of known and novel genes in activated monocytes from patients with rheumatoid arthritis. Arthritis Rheum 2000; 43: 775-90

133. Gerber HP, Vu TH, Ryan AM, et al. VEGF couples hypertrophic cartilage remodeling, ossification and angiogenesis during endochondral bone formation. Nat Med 1999; 5: 623-8

Correspondence and offprints: Dr *Ashok R. Amin*, Rheumatology Research & Laboratory for Functional and Pharmacogenomics, Hospital for Joint Diseases/NYU School of Medicine, Rm 1600, 301 East 17th Street, New York, NY 10003, USA.
E-mail: ashok.amin@msnyuhealth.org

© 2004 Adis Data Information BV. All rights reserved.

View publication stats

NIH Public Access
Author Manuscript

Gastroenterology. Author manuscript; available in PMC 2013 November 15.

Published in final edited form as:
Gastroenterology. 2010 April ; 138(4): . doi:10.1053/j.gastro.2010.02.037.

Genetic Approaches to Functional Gastrointestinal Disorders

YURI A. SAITO[*], **NANDITA MITRA**[‡], and **EMERAN A. MAYER**[§]

[*]C.E.N.T.E.R., Division of Gastroenterology and Hepatology, Mayo Clinic, Rochester, Minnesota

[‡]Department of Biostatistics and Epidemiology, University of Pennsylvania, Philadelphia, Pennsylvania

[§]Center for Neurobiology of Stress, Division of Digestive Diseases, David Geffen School of Medicine at University of California, Los Angeles, Los Angeles, California

Abstract

Functional gastrointestinal disorders are complex symptom-based disorders without agreed upon biomarkers or pathophysiology. A better understanding of the genetic architecture of these disorders would help to better identify their complex biology and explain the common comorbidity with other disorders of persistent pain, mood, and affect, as well as possibly make it possible to identify subgroups of patients who respond to customized therapies. In contrast to monogenic diseases, polygenic diseases and traits are characterized by the contribution of common variants in a large number of genes, as well as environmental factors, to the vulnerability of an individual. Family and twin studies have clearly established a genetic component in irritable bowel syndrome. Although candidate gene studies have identified a few gene polymorphisms that may be correlated with the syndrome, small sample size, lack of reproducibility in large data sets, and the unreliability of the clinical phenotype require caution when extrapolating to a major role of any of the reported polymorphisms in the pathophysiology of irritable bowel syndrome. Future progress in this area will require better characterization of intermediate phenotypes with large effect size for the clinical phenotype, as well as consideration of gene-gene, environment-gene (epigenetics), and sex-gene interactions, genome-wide association, and whole genome sequencing approaches in large data sets.

Keywords

Irritable Bowel Syndrome; Intermediate Phenotypes; Candidate Genes; Genome-wide Association Studies

---

Functional gastrointestinal disorders (FGIDs) are part of a spectrum of complex syndromes that are characterized by persistent pain and discomfort referred to different visceral (irritable bowel syndrome [IBS], functional dyspepsia, interstitial cystitis/painful bladder syndrome) and somatic structures (including fibromyalgia and temporomandibular joint disorder).[1] In the absence of reliable biological markers, or a full understanding of the pathophysiology of any of these syndromes, they are currently defined by symptom criteria and sometimes by the exclusion of “organic” disease.[2] In addition to overlap with each other, these syndromes share several clinical and epidemiologic features, including high

---

© 2010 by the AGA Institute

Address requests for reprints to: Emeran A. Mayer, MD, Center for Neurobiology of Stress, University of California, Los Angeles, 10945 Le Conte Avenue, Suite 2338, Room F, Los Angeles, California 90095-6949. emayer@ucla.edu; fax: (310) 206-3343.

Conflicts of interest
The authors disclose no conflicts.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

comorbidity with symptoms of anxiety and depression, somatization, and stress sensitivity. These similarities have led to the concept that these syndromes, including IBS, share certain pathophysiologic features (eg, central pain amplification) as well as common genetic and environmental vulnerability factors, in addition to having syndrome-specific environmental and genetic factors that are responsible for the individual syndrome manifestation.[1] In the case of FGIDs, such syndrome-specific factors may be related to gastrointestinal function.

There are several fundamental problems when trying to dissect the genetic basis of complex symptom-based disorders, in particular in the case of "functional" disorders that do not have any agreed upon biological marker. The biological mechanisms presenting as symptom-based phenotypes are likely to be heterogeneous. For example, it is likely that certain individuals with mucosal immune activation (eg, lymphocytic colitis; increased mast cell numbers) and those without such changes both meet IBS symptom criteria.[3] In contrast to monogenic diseases, such as cystic fibrosis or Mediterranean fever, in which the variance of the phenotype is completely explained by a single gene, polygenic disorders are determined by many common (and possibly rare) gene variants, each contributing less than 5% and oftentimes much less than 1% of the variance of the clinical phenotype. The common disease/common variant hypothesis states that complex traits result from underlying genetic variants that occur with a relatively high frequency, but the effect size of each of the variants is relatively small.[4,5] A clinical phenotype only emerges if a critical threshold of multiple genetic and environmental factors is reached. The clinical phenotype is not stable. For example, affected patients commonly meet symptom criteria for several of these syndromes at a given time, or at different stages of their lives, and the clinical phenotypes in an individual are not stable over time.[6] Similarly, family histories suggest that relatives and different generations of patients with FGID often do not experience gastrointestinal symptoms but meet criteria of one of the other syndromes. In this review, we will present the current status of our knowledge of the genetic basis of IBS (the FGIDs in which most studies have been conducted), make a critical assessment of these findings, and discuss novel approaches and future directions.

Family and Twin Studies

Familial aggregation has been reproducibly shown in several clinic-based studies of patients with IBS (Table 1), while family studies of other FGIDs are lacking. However, clustering of IBS in families does not necessarily implicate a genetic basis, but could also result from environmental exposures shared in a household, including diet or lifestyle behaviors, exposure to adverse life events within the family, learned cognitions about disease and illness behavior, or even shared exposure to microorganisms.[7,8]

Twin studies are the traditional method used to estimate the genetic and environmental contributions to disease. To date, 5 twin studies of FGIDs, mostly of IBS, have been published (Table 2). Based on these studies, the genetic liability of IBS and FGIDs ranges between 0 and 20% based on concordance rates, and the genetic heritability of FGIDs was estimated to be between 22% and 57%. The variation in heritability estimates likely is due to differences in study methodology. Importantly, in these studies, general shared environment was not a significant predictor of phenotypic variance. However, in the Virginia Twin Study, it was also observed that having a mother or father with IBS was an independent predictor of IBS status, which was stronger than having a twin with IBS, suggesting that social learning or stability of the family environment (which might be influenced by anxiety or depression in one of the parents, disorders often associated with IBS) may have greater influence on the development of IBS than genetics alone.[9]

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Candidate Gene Studies

Candidate gene studies are studies whereby investigators postulate that a specific gene, and typically a specific functional gene polymorphism that results in alterations in protein function or quantity, may play a role in disease pathophysiology. The most commonly used approach in IBS to date has been to search for correlations of a candidate gene polymorphism with the symptom-based phenotype IBS (Table 3). More recently, a few studies have aimed to identify correlations between gene polymorphism and biological intermediate phenotypes (see the following text), such as colonic transit, enhanced perception of visceral stimuli, or altered brain responses.[10,11]

Associations of Gene Polymorphisms With Clinical Phenotype

As shown in Table 3, more than a dozen candidate genes have been studied for possible associations with the clinical phenotype of IBS. The majority of genes studied encode serotonin-related proteins (eg, serotonin trans-porter, 5-HT$_3$ receptor, 5-HT$_2$ receptor), proteins involved in noradrenergic signaling (eg, α-adrenergic receptor), and genes encoding immune markers (eg, interleukin-10, transforming growth factor β1). Others genes include the β3 subunit of the G protein (*GNβ3*), cholecystokinin type 1 receptor (*CCK1-R*), Nav1.5 sodium channel (*SCN5A*), and other intestinal epithelial barrier factors.

The best-studied polymorphisms are those related to serotonin (5-hydroxytryptamine [5-HT]) signaling in the brain-gut axis, including the serotonin transporter (SERT), and several 5-HT receptors. This emphasis on genetic variations within the 5-HT signaling system is based on the excellent construct validity of this system from both preclinical[12] and clinical[13] studies. Parallel alterations in central and gut-related 5-HT signaling could explain the typical constellation of altered motility and secretion associated with increased stress responsiveness and increased anxiety. The 5-HTT LPR polymorphism (for serotonin transporter–linked polymorphic region) in the promoter region of the serotonin transporter gene *SLC6A4* is the most intensively studied polymorphism in IBS and has also been studied in psychiatric syndromes,[14] fibromyalgia, and temporomandibular joint disorder.[15–17] Surprisingly, several small case-control studies performed in patients with IBS have shown variable results, and a recent meta-analysis suggests that there is no simple association between this genetic variant and the clinical phenotype of IBS.[18] There has been one study showing a potential association between another functional single nucleotide polymorphism (SNP) (rs25531) on the *SLC6A4* gene with IBS and another study linking the 76G>A SNP on the 5-HT$_3$ receptor type E subunit gene (replicated in a second study) and the 489C>A SNP on the *HTR3C* gene with diarrhea-predominant IBS (IBS-D).

The majority of the other polymorphisms shown in Table 3 have not been found to be associated with the IBS clinical phenotype. However, the majority of these candidate gene/clinical phenotype association studies were small (<200 subjects), did not take into account the caveats addressed in the following section, and with the exception of the 5-HTT LPR and the 5-HT$_{3E}$ gene polymorphism studies have never been replicated in a separate cohort. Due to these limitations, very few of these reported findings allow us to make any conclusions about their validity and, in the opinion of the authors, should not be implicated in the pathophysiology of IBS at this point.

Association of Candidate Gene With Intermediate Phenotype

Intermediate phenotypes, also known as endophenotypes, are biologically related to the clinical phenotype of interest, stable over time and reproducible in different laboratories, detectable at baseline or in the presence of a provocative stimulus, and often present in unaffected family members.[19,20] In contrast to the clinical phenotype, which is based on

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

subjective and unstable patient reports (see the previous text), intermediate phenotypes are biological or physiologic variables that can be objectively measured and are believed to be more closely related to genetic factors (see Figure 1). This candidate genotype/intermediate phenotype approach has been successfully applied to many polygenic disorders, including chronic pain and schizophrenia.[21] For example, polymorphisms of the *COMT* gene have been associated with objective measures of pain sensitivity[17] and with endogenous opioid release in the brain.[22] Unfortunately, no intermediate endophenotype has been described to date for FGIDs that meets all the criteria previously listed.

Camilleri et al pioneered the intermediate endophenotype approaches in gastrointestinal disorders.[10,11] For example, in one study, several significant gene-sensory and gene-motor associations were observed in IBS and healthy subjects involving several adrenergic and serotonergic gene polymorphisms.[23] Fukudo et al studied the relationship between 5-HTT LPR genotype and regional cerebral blood flow using positron emission tomography during and without colorectal balloon inflation, with the idea that altered brain activation by rectal stimulation might serve as an IBS endophenotype. The 5-HTT LPR polymorphism s/s genotype appeared to be associated with altered brain activation[24,25] and with connectivity within an emotional arousal network with disinhibition of the amygdala,[26] suggesting that this area of the brain may mediate central pain amplification seen in patients with IBS.

Limitations and Interpretation of Current Candidate Gene Studies

False-positive findings—The great majority of findings from small samples (such as those used in most FGID studies) represent false positives that cannot be replicated in a second sample. Because of the large number of genes (20,000) and the even larger number of testable genetic variants (millions), investigators will observe positive associations between hypothesized polymorphisms and complex FGID by chance. Many advocate the use of a higher statistical threshold for significance if multiple genetic variants are tested and/or comparisons (eg, cases vs controls, case-subtype A vs controls, case-subtype B vs controls, and so on) are performed. For example, if 100 SNPs are tested, then the P value to be considered statistically significant after a Bonferroni correction would be .0005. Such high thresholds require either a large genetic effect whereby the variant is very common in cases but relatively uncommon in controls (highly unlikely for FGIDs) and/or very large sample sizes (thousands of patients). Less conservative corrections for multiple comparisons such as the Benjamini–Hochberg approach to controlling the false discovery rate[26] are also available and appropriate but do not entirely alleviate the multiple-testing burden. In addition to sample size, replication of positive studies (lacking for most FGID studies) is critically important, if not an absolute requirement to rule out type I and II errors. Although candidate gene association studies are attractive to test specific hypotheses related to the pathophysiology of FGIDs, the results of one well-designed study still require additional field-based and laboratory-based corroboration before firmly concluding that a genetic variant truly lies in the causal pathway of disease or disorder development.

Disease phenotype—Most studies to date have used conventional symptom-based criteria for defining the presence and type of FGID. The problem with such an approach is that it is dichotomous (eg, affected or unaffected), with no accommodation for intermediate states that are common in polygenic disorders. Similar to many syndromes in psychiatry, such as depression and attention-deficit disorders, such categorical disease definitions are probably artificial because they do not take into account (1) the continuum of symptoms from healthy individuals to health care–seeking patients; (2) the spectrum and overlap of FGIDs with each other as well as other disorders (eg, depression, anxiety, somatization, fibromyalgia); (3) the lack of stability of symptoms over time if disease definition is based on symptom reporting for a set period (eg, "last year," "past 3 months"); and (4) the

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

underlying biological mechanisms, grouping people with similar symptoms but different mechanisms for symptom experience or reporting.

Small sample size—Most studies (including all the studies listed in Table 3) do not have sufficiently large sample sizes to address whether a genetic variant may be associated with a specific disorder subtype. It is biologically plausible that a genetic variant may not be causally relevant for the overall disorder but may be responsible for a specific disorder subtype or symptom, particularly if there is clinical and presumed underlying biological heterogeneity. For example, the 5-HTT LPR s allele is not associated with IBS as a group but may be associated with IBS-D or with patients with IBS with high comorbid anxiety. Furthermore, the magnitude of effect of genetic contributors for IBS and other FGIDs is likely quite modest. If at best a few genes contribute about 5% of the variance and the majority less than 1%, and the clinical phenotypes are unreliable (see the introduction), a very large number of subjects (several thousand) are required. In addition, identification of susceptibility loci for FGIDs will likely require a genome-wide approach, testing thousands if not millions of SNPs, whereby multiple-testing issues apply. Finally, analyses studying multiple interactions (eg, gene 1 × gene 2 × environment × sex) will also need to be enlarged to take into account additional variables.

Control groups—Some genetic studies use “convenience” samples of DNA collected from other studies without information about disease/disorder status. Contamination by misclassification may be particularly important for common symptom-based disorders such as FGIDs. Ideally, control subjects should not have a present or past history of an FGID or other comorbid somatic or psychiatric disorders, or a family history of such disorders, and should be recruited prospectively in a similar manner to case recruitment. Also, because functional symptoms present in adulthood and are female predominant, and because genetic variation may differ by race, cases and controls should ideally be matched by age, sex, and race. For example, confounding due to population stratification arises when differences in both allele and disease frequencies exist in a population of mixed racial/ethnic subpopulations. The presence of population stratification could potentially severely bias genetic effect estimates, leading to both false-negative and false-positive results. Most commonly used approaches for adjusting for population stratification include principal components analysis[27] and multidimensional scaling.[28]

Lack of gene-gene, gene-environment, and gene-sex interactions—FGIDs are complex polygenic disorders with significant sex and environmental effects. Prior epidemiologic studies and the previously cited twin studies[9] support this concept, and an extensive literature supports the importance of epigenetic factors in determining the clinical phenotype. For example, it has been shown that unless the environmental factor of adverse life events is taken into account, there is no relationship between the 5-HTT LPR polymorphism and the clinical trait of depression.[29] The effect of adverse early life events influences the adult phenotype through gene modification, such as methylation,[30] and the phenotype of SERT knockout mice varies with the sex of the animal.[31]

Future Directions

Genome-wide Association Studies and Whole Genome Sequencing

More than 100 genome-wide association studies have been conducted for a large number of human diseases, identifying hundreds of polymorphisms with influence on disease vulnerability.[32] However, in most of these studies, one or two genes that had already been known were confirmed, while a large number of common gene variants found in these studies are responsible only for a small fraction of the genetic variation that we know from

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

family and twin studies. It is likely that a large number of these genes are connected by pathways forming extensive networks with each other and with the various phenotypes. In FGIDs, this approach is further complicated by the fact that in contrast to many of the disease studies with genome-wide association studies (including inflammatory bowel disease), there are few biological markers (such as gastrointestinal transit, gene expression profiles) or validated intermediate phenotypes for FGIDs. Nevertheless, with the anticipated continuing decrease in the costs of performing such genome-wide studies (genome-wide association studies and whole genome sequencing), combined with a consortium approach to study a large number of well-phenotyped patients (as is under way in other complex diseases such as Alzheimer's disease or major psychiatric diseases), such approaches are likely to be applied to FGIDs and related persistent pain disorders in the near future. Systems biological approaches will be required to characterize the complex interactions between hundreds of genes and multiple intermediate phenotypes.[33]

Better Phenotyping

With the rapid advances in large-scale genotyping approaches, the bottleneck in the better understanding of the genetic underpinnings of complex diseases such as FGIDs is the identification and validation of robust intermediate phenotypes (see the previous text). This will require a shift of emphasis from categorical syndrome definition of individual FGIDs ("agnostic approach") and of individual persistent pain disorders to a view of these syndromes as a "mosaic" made up of a large number of interrelated phenotypes.[34] Such correlations between intermediate phenotypes have been reported in a variety of diseases, including disorders of metabolism[35] and schizo-phrenia,[36] and the approach has been adopted in FGID studies by some investigators (see Association of Candidate Genes With Intermediate Phenotype). Among many highly correlated phenotypes, the one with the highest reliability and effect size for the clinical syndrome should be optimal to use for detecting associations with particular genotypes.[34]

Novel Analytic Approaches

As genotyping technology rapidly advances and improves, novel analytic methods need to be developed and appropriately applied. Single SNP analyses using χ^2 tests, Armitage trend tests, and multivariable logistic regression are often sufficient to gain a first impression of genetic associations; however, haplotype-based[37,38] and gene-based and pathway-based methods[39 – 44] often allow a more comprehensive and integrated view of the network of associations and interactions that may exist. Importantly, focus has recently been shifting to the discovery of rare variants via sequencing, and attention should be given to optimal strategies for analyzing next-generation sequencing data.[45– 47]

It is also suggested that some thought be given to whether a purely frequentist approach for assessing evidence for a genetic association should be adopted or whether a Bayesian approach would be more informative. In the frequentist approach, a *P* value is computed for the null hypothesis of no association. However, a *P* value alone does not fully quantify how confident we should be that the variant is truly associated with the phenotype. Bayesian methodology provides an alternative approach that overcomes the limitations of the *P* value (but with the added cost of additional modeling assumptions). The Bayesian approach provides a measure of evidence (the Bayes factor) that allows for a practical and quantitative way of incorporating biological information in the association estimate, allows the direct comparison of SNPs across studies, and provides a comprehensive approach to combining results over several studies in a meta-analysis and across genes in pathway-based analysis.[48] New and innovative statistical methods are rapidly being developed in the hope of facilitating more robust conclusions and interpretable results in future genetic association

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

studies. Investigators are advised to stay abreast of these statistical advances and to adopt them into their studies.

Summary and Conclusions

Gene discovery for complex polygenic disorders such as FGIDs remains a priority but represents a vast challenge. The field, in particular the individual candidate gene hunting approach, has been hampered by use of discrete and narrow symptom-based definitions—focusing on one disorder such as IBS—instead of objective biological phenotypes. Future studies seeking to identify causal genetic loci for FGIDs will likely use a combination of approaches, with candidate gene as well as genome-wide approaches. Whole genome sequencing in combination with expression profiling and novel analysis techniques will not only enable researchers to identify variations in individual genes, but also the effect of epigenetic processes such as gene methylation on gene expression, reflecting environmental influences (Figure 2).

Acknowledgments

Funding

Supported by a grant from the National Institutes of Health (DK76797 to Y.A.S.) and (DK48351, DK64539, and AT002681 to E.A.M.).

Abbreviations used in this paper

FGID	functional gastrointestinal disorder
5-HT	5-hydroxytryptamine
IBS	irritable bowel syndrome
IBS-D	diarrhea-predominant irritable bowel syndrome
SNP	single nucleotide polymorphism

References

1. Mayer, EA.; Bushnell, MC. Functional pain disorders: time for a paradigm shift?. In: Mayer, EA.; Bushnell, MC., editors. Functional pain syndromes: presentation and pathophysiology. Seattle, WA: IASP Press; 2009. p. 531-565.
2. Longstreth GF, Thompson WG, Chey WD, et al. Functional bowel disorders. Gastroenterology. 2006; 130:1480–1491. [PubMed: 16678561]
3. Chadwick VS, Chen W, Shu D, et al. Activation of the mucosal immune system in irritable bowel syndrome. Gastroenterology. 2002; 122:1778–1783. [PubMed: 12055584]
4. Diatchenko L, Nackley AG, Slade GD, et al. Idiopathic pain disorders—pathways of vulnerability. Pain. 2006; 123:226–230. [PubMed: 16777329]
5. Diatchenko L, Slade GD, Nackley AG, et al. Genetic basis for individual variations in pain perception and the development of chronic pain condition. Hum Mol Genet. 2005; 14:135–143. [PubMed: 15537663]
6. Halder SL, Locke GR III, Schleck CD, et al. Natural history of functional gastrointestinal disorders: a 12-year longitudinal population-based study. Gastroenterology. 2007; 133:799–807. [PubMed: 17678917]
7. Kalantar JS, Locke GR III, Zinsmeister AR, et al. Familial aggregation of irritable bowel syndrome: a prospective study. Gut. 2003; 52:1703–1707. [PubMed: 14633946]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

8. Saito YA, Zimmerman JM, Harmsen WS, et al. Irritable bowel syndrome aggregates strongly in families: a family-based case-control study. Neurogastroenterol Motil. 2008; 20:790–797. [PubMed: 18221250]
9. Levy RL, Jones KR, Whitehead WE, et al. Irritable bowel syndrome in twins: heredity and social learning both contribute to etiology. Gastroenterology. 2001; 121:799–804. [PubMed: 11606493]
10. Camilleri M, Carlson P, Zinsmeister AR, et al. Neuropeptide s receptor induces neuropeptide expression and associates with intermediate phenotypes of functional gastrointestinal disorders. Gastroenterology. 2010; 138:98–107. [PubMed: 19732772]
11. Papathanasopoulos A, Camilleri M, Carlson PJ, et al. A preliminary candidate genotype-intermediate phenotype study of satiation and gastric motor function in obesity. Obesity (Silver Spring). 2009 Oct 29. [Epub ahead of print].
12. Murphy DL, Lesch KP. Targeting the murine serotonin transporter: insights into human neurobiology. Nat Rev Neurosci. 2008; 9:85–96. [PubMed: 18209729]
13. Gershon MD, Tack J. The serotonin signaling system: from basic understanding to drug development for functional GI disorders. Gastroenterology. 2007; 132:397–414. [PubMed: 17241888]
14. Hariri AR, Holmes A. Genetics of emotional regulation: the role of the serotonin transporter in neural function. Trends Cogn Sci. 2006; 10:182–191. [PubMed: 16530463]
15. Cohen H, Buskila D, Neumann L, et al. Confirmation of an association between fibromyalgia and serotonin transporter promoter region (5-HTTLPR) polymorphism, and relationship to anxiety-related personality traits. Arthritis Rheum. 2002; 46:845– 847. [PubMed: 11920428]
16. Offenbaecher M, Bondy B, de Jonge S, et al. Possible association of fibromyalgia with a polymorphism in the serotonin transporter gene regulatory region. Arthritis Rheum. 1999; 42:2482–2488. [PubMed: 10555044]
17. Diatchenko L, Nackley AG, Slade GD, et al. Catechol-O-methyl-transferase gene polymorphisms are associated with multiple pain-evoking stimuli. Pain. 2006; 125:216–224. [PubMed: 16837133]
18. Van Kerkhoven LA, Laheij RJ, Jansen JB. Meta-analysis: a functional polymorphism in the gene encoding for activity of the serotonin transporter protein is not associated with the irritable bowel syndrome. Aliment Pharmacol Ther. 2007; 26:979–986. [PubMed: 17877505]
19. Gottesman, Gould TD. The endophenotype concept in psychiatry: etymology and strategic intentions. Am J Psychiatry. 2003; 160:636–645. [PubMed: 12668349]
20. Mayer EA. The challenge of studying the biology of complex, symptom-based GI disorders. Gastroenterology. 2008; 134:1826–1827. [PubMed: 18477481]
21. Meyer-Lindenberg A, Weinberger DR. Intermediate phenotypes and genetic mechanisms of psychiatric disorders. Nat Rev Neurosci. 2006; 7:818–827. [PubMed: 16988657]
22. Zubieta JK, Heitzeg MM, Smith YR, et al. COMT val 158 met genotype affects mu-opioid neurotransmitter responses to a pain stressor. Science. 2003; 299:1240–1243. [PubMed: 12595695]
23. Camilleri M, Busciglio I, Carlson P, et al. Candidate adrenergic, GNB3, and serotonergic polymorphisms, endophenotype, pharmacogenetics of clonidine in irritable bowel syndrome and health. Gastroenterology. 2008
24. Labus JS, Mayer EA, Hamaguchi T, et al. 5-Httlpr gene polymorphism modulates activity and connectivity within an emotional arousal network of healthy control subjects during visceral pain (abstr). Gastroenterology. 2008; 134:A121.
25. Fukudo S, Kanazawa M, Mizuno T, et al. Impact of serotonin transporter gene polymorphism on brain activation by colorectal distention. Neuroimage. 2009; 47:946–951. [PubMed: 19426812]
26. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B (Methodological). 1995; 57:12.
27. Price AL, Patterson NJ, Plenge RM, et al. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006; 38:904–909. [PubMed: 16862161]
28. Li Q, Yu K. Improved correction for population stratification in genome-wide association studies by identifying hidden population structures. Genet Epidemiol. 2008; 32:215–226. [PubMed: 18161052]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

29. Caspi A, Sugden K, Moffitt TE, et al. Influence of life stress on depression: Moderation by a polymorphism in the 5-HTT gene. Science. 2003; 301:386–389. [PubMed: 12869766]
30. Fish EW, Shahrokh D, Bagot R, et al. Epigenetic programming of stress responses through variations in maternal care. Ann N Y Acad Sci. 2004; 1036:167–180. [PubMed: 15817737]
31. Cornelissen LL, Brooks DP, Wibberley A. Female, but not male, serotonin reuptake transporter (5-HTT) knockout mice exhibit bladder instability. Auton Neurosci. 2005; 122:107–110. [PubMed: 16023897]
32. Goldstein DB. Common genetic variation and human traits. N Engl J Med. 2009; 360:1696–1698. [PubMed: 19369660]
33. Papp, J.; Sobel, E. The genetics of pain. In: Mayer, EA.; Bushnell, MC., editors. Functional pain syndromes: presentation and pathophysiology. Seattle, WA: IASP Press; 2009. p. 405-422.
34. Ioannidis JP, Thomas G, Daly MJ. Validating, augmenting and refining genome-wide association signals. Nat Rev Genet. 2009; 10:318–328. [PubMed: 19373277]
35. Martin LJ, North KE, Dyer T, et al. Phenotypic, genetic, and genome-wide structure in the metabolic syndrome. BMC Genet. 2003; 4(Suppl 1):S95. [PubMed: 14975163]
36. Aukes MF, Alizadeh BZ, Sitskoorn MM, et al. Genetic overlap among intelligence and other candidate endophenotypes for schizophrenia. Biol Psychiatry. 2009; 65:527–534. [PubMed: 19013556]
37. Schaid DJ, Rowland CM, Tines DE, et al. Score tests for association between traits and haplotypes when linkage phase is ambiguous. Am J Hum Genet. 2002; 70:425–434. [PubMed: 11791212]
38. Lake SL, Lyon H, Tantisira K, et al. Estimation and tests of haplotype-environment interaction when linkage phase is ambiguous. Hum Hered. 2003; 55:56–65. [PubMed: 12890927]
39. Conti DV, Cortessis V, Molitor J, et al. Bayesian modeling of complex metabolic pathways. Hum Hered. 2003; 56:83–93. [PubMed: 14614242]
40. Kooperberg C, Ruczinski I. Identifying interacting SNPs using Monte Carlo logic regression. Genet Epidemiol. 2005; 28:157–170. [PubMed: 15532037]
41. Satagopan JM, Yandell BS, Newton MA, et al. A bayesian approach to detect quantitative trait loci using Markov chain Monte Carlo. Genetics. 1996; 144:805–816. [PubMed: 8889541]
42. Kesh S, Mensah NY, Peterlongo P, et al. TLR1 and TLR6 polymorphisms are associated with susceptibility to invasive aspergillosis after allogeneic stem cell transplantation. Ann N Y Acad Sci. 2005; 1062:95–103. [PubMed: 16461792]
43. Ritchie MD, Hahn LW, Roodi N, et al. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am J Hum Genet. 2001; 69:138–147. [PubMed: 11404819]
44. Chatterjee N, Kalaylioglu Z, Moslehi R, et al. Powerful multilocus tests of genetic association in the presence of gene-gene and gene-environment interactions. Am J Hum Genet. 2006; 79:1002–1016. [PubMed: 17186459]
45. Rusk N. Focus on next-generation sequencing data analysis. Foreword Nat Methods. 2009; 6:S1.
46. Medvedev P, Stanciu M, Brudno M. Computational methods for discovering structural variation with next-generation sequencing. Nat Methods. 2009; 6:S13–S20. [PubMed: 19844226]
47. Li B, Leal SM. Discovery of rare variants via sequencing: implications for the design of complex trait association studies. PLoS Genet. 2009; 5:e1000481. [PubMed: 19436704]
48. Stephens M, Balding DJ. Bayesian statistical methods for genetic association studies. Nat Rev Genet. 2009; 10:681–690. [PubMed: 19763151]
49. Whorwell PJ, McCallum M, Creed FH, et al. Non-colonic features of irritable bowel syndrome. Gut. 1986; 27:37–40. [PubMed: 3949235]
50. Bellentani S, Baldoni P, Petrella S, et al. A simple score for the identification of patients at high risk of organic diseases of the colon in the family doctor consulting room. The Local IBS Study Group. Fam Pract. 1990; 7:307–312. [PubMed: 2289644]
51. Locke GR III, Zinsmeister AR, Talley NJ, et al. Familial association in adults with functional gastrointestinal disorders. Mayo Clin Proc. 2000; 75:907–912. [PubMed: 10994826]
52. Levy RL, Whitehead WE, Von Korff MR, et al. Intergenerational transmission of gastrointestinal illness behavior. Am J Gastroenterol. 2000; 95:451–456. [PubMed: 10685749]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

53. Kanazawa M, Endo Y, Whitehead WE, et al. Patients and nonconsulters with irritable bowel syndrome reporting a parental history of bowel problems have more impaired psychological distress. Dig Dis Sci. 2004; 49:1046–1053. [PubMed: 15309899]
54. Pata C, Erdal ME, Derici E, et al. Serotonin transporter gene polymorphism in irritable bowel syndrome. Am J Gastroenterol. 2002; 97:1780–1784. [PubMed: 12135035]
55. Kim HJ, Camilleri M, Carlson PJ, et al. Association of distinct alpha(2) adrenoceptor and serotonin transporter polymorphisms with constipation and somatic symptoms in functional gastrointestinal disorders. Gut. 2004; 53:829–837. [PubMed: 15138209]
56. Yeo A, Boyd P, Lumsden S, et al. Association between a functional polymorphism in the serotonin transporter gene and diarrhoea predominant irritable bowel syndrome in women. Gut. 2004; 53:1452–1458. [PubMed: 15361494]
57. Lee DY, Park H, Kim WH, et al. Serotonin transporter gene polymorphism in healthy adults and patients with irritable bowel syndrome [in Korean]. Korean J Gastroenterol. 2004; 43:18–22. [PubMed: 14745247]
58. Park JM, Choi MG, Park JA, et al. Serotonin transporter gene polymorphism and irritable bowel syndrome. Neurogastroenterol Motil. 2006; 18:995–1000. [PubMed: 17040410]
59. Saito YA, Locke GR III, Zimmerman JM, et al. A genetic association study of 5-HTT LPR and GNbeta3 C825T polymorphisms with irritable bowel syndrome. Neurogastroenterol Motil. 2007; 19:465–470. [PubMed: 17564628]
60. Li Y, Nie Y, Xie J, et al. The association of serotonin transporter genetic polymorphisms and irritable bowel syndrome and its influence on tegaserod treatment in Chinese patients. Dig Dis Sci. 2007; 52:2942–2949. [PubMed: 17394071]
61. Sikander A, Rana S, Sinha S, et al. Serotonin transporter promoter variant: Analysis in Indian IBS patients and control population. J Clin Gastroenterol. 2009; 43:902–904. [PubMed: 19727006]
62. Niesler B, Kapeller J, Fell C, et al. 5-HTTLPR and STin2 polymorphisms in the serotonin transporter gene and irritable bowel syndrome: effect of bowel habit and sex. Eur J Gastroenterol Hepatol. 2009 Jun 25. [Epub ahead of print].
63. Kohen R, Jarrett ME, Cain KC, et al. The serotonin transporter polymorphism rs25531 is associated with irritable bowel syndrome. Dig Dis Sci. 2009 Jan 1. [Epub ahead of print].
64. Saito Y, Larson J, Atkinson E, et al. The role of 5-HTT LPR and GNbeta3 825C>T polymorphisms and gene environment interactions in irritable bowel syndrome. Gastroenterology. 2009; 136:289.
65. Pata C, Erdal E, Yazc K, et al. Association of the -1438 G/A and 102 T/C polymorphism of the 5-Ht2A receptor gene with irritable bowel syndrome 5-Ht2A gene polymorphism in irritable bowel syndrome. J Clin Gastroenterol. 2004; 38:561–566. [PubMed: 15232358]
66. Kapeller J, Houghton L, Monnikes H, et al. First evidence for an association of functional variant in the microRNA-510 target site of the serotonin receptor-type 3E gene with diarrhea predominant irritable bowel syndrome. Hum Mol Genet. 2008; 17:2967–2977. [PubMed: 18614545]
67. Kapeller J, Houghton L, Walstab J, et al. A coding variant in the serotonin receptor 3C subunit is associated with diarrhea-predominant irritable bowel syndrome (abstr). Gastroenterology. 2009; 136:A155–A156.
68. Gonsalkorale WM, Perrey C, Pravica V, et al. Interleukin 10 genotypes in irritable bowel syndrome: evidence for an inflammatory component? Gut. 2003; 52:91–93. [PubMed: 12477767]
69. van der Veek P, de Kroon Y, Verspaget H, et al. Tumor necrosis factor alpha and interleukin 10 gene polymorphisms in irritable bowel syndrome (abstr). Gastroenterology. 2004; 126:A53.
70. Sikander A, Rana SV, Prasad KK. Role of serotonin in gastrointestinal motility and irritable bowel syndrome. Clin Chim Acta. 2009; 403:47–55. [PubMed: 19361459]
71. Holtmann G, Siffert W, Haag S, et al. G-protein beta 3 subunit 825 CC genotype is associated with unexplained (functional) dyspepsia. Gastroenterology. 2004; 126:971–979. [PubMed: 15057736]
72. Andresen V, Camilleri M, Kim HJ, et al. Is there an association between GNbeta3-C825T genotype and lower functional gastrointestinal disorders? Gastroenterology. 2006; 130:1985–1994. [PubMed: 16762621]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

73. Colucci R, Ghisu N, Bellini M, et al. Analysis of cholecystokinin type 1 receptor gene polymorphisms in patients with irritable bowel syndrome (abstr). Gastroenterology. 2009; 136:A225.

74. Saito YA, Strege PR, Tester DJ, et al. Sodium channel mutation in irritable bowel syndrome: evidence for an ion channelopathy. Am J Physiol Gastrointest Liver Physiol. 2009; 296:G211–G218. [PubMed: 19056759]

75. Villani A, Saito Y, Lemire M, et al. Validation of genetic risk factors for post-infectious irritable bowel syndrome (IBS) in patients with sporadic IBS. Gastroenterology. 2009; 136:289.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

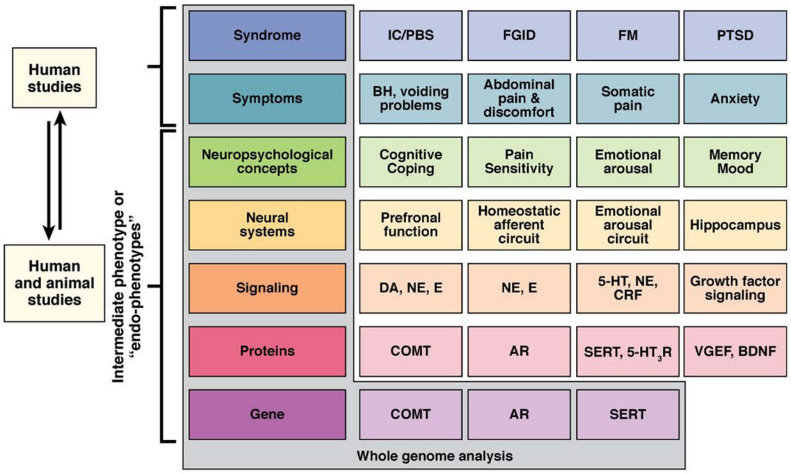

Human studies
Human and animal studies
Intermediate phenotype or "endo-phenotypes"
Syndrome
IC/PBS
FGID
FM
PTSD
Symptoms
BH, voiding problems
Abdominal pain & discomfort
Somatic pain
Anxiety
Neuropsychological concepts
Cognitive Coping
Pain Sensitivity
Emotional arousal
Memory Mood
Neural systems
Prefronal function
Homeostatic afferent circuit
Emotional arousal circuit
Hippocampus
Signaling
DA, NE, E
NE, E
5-HT, NE, CRF
Growth factor signaling
Proteins
COMT
AR
SERT, 5-HT$_3$R
VGEF, BDNF
Gene
COMT
AR
SERT
Whole genome analysis

Figure 1.

Intermediate phenotype strategy for FGIDs and frequently comorbid disorders. Shown is the progressive deconstruction of clinical, symptom-based syndromes (several examples are shown in the first row) into intermediate phenotypes (examples are shown in each row) all the way down to the genome. For clarity, only a small number of candidate genes are shown. Although the syndrome and the symptoms can only be studied in humans, the neurobiological endophenotypes can be studied both in humans and in animal models. It has been postulated that the correlation between neurobiological endophenotypes and genes is significantly higher than the correlation between clinical syndromes and genes. IC/PBS, interstitial cystitis/painful bladder syndrome; FM, fibromyalgia; PTSD, posttraumatic stress syndrome; DA, dopamine; NE, norepinephrine; E, epinephrine; CRF, corticotropin-releasing factor; COMT, catechol-*O*-methyltransferase; AR, adrenoreceptor; VGEF, vascular endothelial growth factor; BDNF, brain-derived neurotrophic factor. Mayer EA, Bushnell MC. Functional pain disorders: time for a paradigm shift? In: Mayer EA, Bushnell MC, eds. Functional pain syndromes: presentation and pathophysiology. Seattle, WA: IASP Press, 2009:531–565. This figure has been reproduced with permission of the International Association for the Study of Pain® (IASP®). The figure may not be reproduced for any other purpose without permission.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

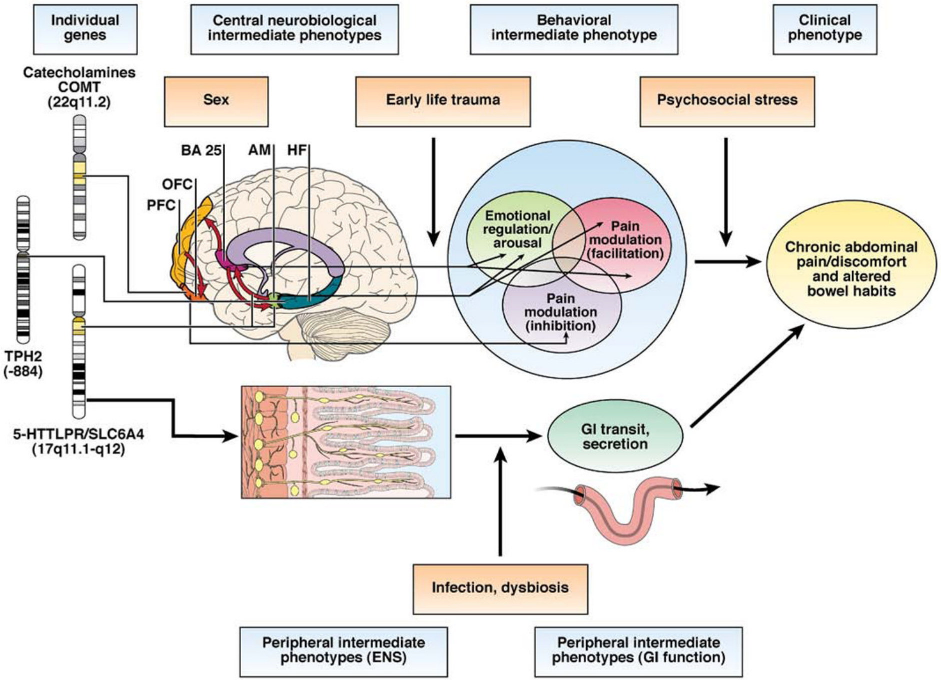

Individual genes
Central neurobiological intermediate phenotypes
Behavioral intermediate phenotype
Clinical phenotype
Catecholamines COMT (22q11.2)
Sex
Early life trauma
Psychosocial stress
BA 25
AM
HF
OFC
PFC
Emotional regulation/ arousal
Pain modulation (facilitation)
Pain modulation (inhibition)
Chronic abdominal pain/discomfort and altered bowel habits
TPH2 (-884)
5-HTTLPR/SLC6A4 (17q11.1-q12)
GI transit, secretion
Infection, dysbiosis
Peripheral intermediate phenotypes (ENS)
Peripheral intermediate phenotypes (GI function)

Figure 2.

Simplified model for interaction between genes, peripheral and central endophenotypes, and environmental factors resulting in IBS symptoms. Individual genes influence the function of neural networks in the brain and in the enteric nervous system (including closely related gut cells). Alterations in these neural systems result in alterations in central and peripheral systems related to pain perception, autonomic arousal, peristalsis, and secretion. Significant interactions of genetic with epigenetic factors (including microflora environmental stressors) and with the sex of the individual are likely. PFC, prefrontal cortex; OFC, orbitofrontal cortex; BA 25, Brodman area 25; AM, amygdala. TPH2, tryptophan hydroxylase 2. Mayer EA, Bushnell MC. Functional pain disorders: time for a paradigm shift? In: Mayer EA, Bushnell MC, eds. Functional pain syndromes: presentation and pathophysiology. Seattle, WA: IASP Press, 2009:531–565. This figure has been reproduced with permission of the International Association for the Study of Pain® (IASP®). The figure may not be reproduced for any other purpose without permission.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 1

Family Studies of IBS

Citation	Finding
Whorwell et al, *Gut* 1986[49]	33% of patients with IBS vs 2% of controls have a family history of IBS
Bellentani et al, *Fam Pract* 1990[50]	IBS in a first-degree relative predicts a diagnosis of IBS in patients with abdominal pain
Locke et al, *Mayo Clin Proc* 2000[51]	Reporting a relative with abdominal pain or altered bowel habit predicts IBS (odds ratio, 2.5; 95% confidence interval, 1.5–4.2)
Levy et al, *Am J Gastroenterol* 2000[52]	Health maintenance organization database study: children of parents with IBS are more likely to present to providers with abdominal pain than children of parents without IBS
Kanazawa et al, *Dig Dis Sci* 2004[53]	Familial aggregation of IBS is not due to the result of psychological factors or psychiatric disorders
Saito et al, *Neurogastroenterol Motil* 2008[8]	Case relatives are 3 times more likely to meet Rome criteria for IBS than control relatives by direct survey

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 2

Twin Studies of IBS

	No. of twin pairs	Case definition	Data collection method	Genetic liability (%)	Heritability in best-fit model (%)		
					Additive genetic	**Shared environment**	**Individual environment**
Australian Twin Study	343	Functional bowel disease	No. of symptoms reported by interview	20	57	0	43
Virginia Twin Study	6060	IBS	Check box by IBS	8.8	—	—	—
Minnesota Twin Study	986	IBS by Rome II	Validated bowel disease questionnaire	16	22	0	78
Norwegian Twin Study	3334	IBS	Check box by IBS	13.3	48	0	52
British Twin Study	1870	IBS by Rome II	Validated questionnaire	1	0	26	74

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

Table 3

Candidate Gene Studies in IBS

Protein (gene)	Polymorphism	Reference	Case sample size	Finding summary
Serotonin pathway				
Serotonin transporter (*SLC6A4*)	5-HTT LPR	Pata et al, *Am J Gastroenterol* 2002[54]	54	No association; LS associated with IBS-D
		Kim et al, *Gut* 2004[55]	256	No association
		Yeo et al, *Gut* 2004[56]	186	Only patients with IBS-D; SS associated with IBS-D
		Lee et al, *Korean J Gastroenterol* 2004[57]	33	No association
		Park et al, *Neurogastroenterol Motil* 2006[58]	190	
		Saito et al, *Neurogastroenterol Motil* 2007[59]	50	No association; S associated with IBS with a mixed bowel pattern
		Li et al, *Dig Dis Sci* 2007[60]	87	No association; LL associated with IBS-C
		Sikander et al, *J Clin Gastroenterol* 2009[61]	151	No association; SS associated with IBS-C
		Niesler et al, *Eur J Gastroenterol Hepatol* 2009[62]	196	N/A; no association with IBS-C or IBS-D; lower frequency of SS in male IBS-D
		Kohn et al, *Dig Dis Sci* 2009[63]	186	No association with IBS
		Saito et al, *Gastroenterology* 2009[64]	312	No association with IBS or IBS subtypes
	STin2 VNTR	Pata et al, *Am J Gastroenterol* 2002[54]	54	No association with IBS or IBS subtypes
		Niesler et al, *Eur J Gastroenterol Hepatol* 2009[62]	196	No overall IBS analysis performed; no association with IBS-C or IBS-D
		Kohn et al, *Dig Dis Sci* 2009[63]	186	No association
	rs25531	Kohn et al, *Dig Dis Sci* 2009[63]	186	G allele associated with IBS
5-HT$_{2A}$ receptor subunit (HTR2A)	102T>C	Pata et al, *J Clin Gastroenterol* 2004[65]	54	No association
	−1438G>A	Pata et al, *J Clin Gastroenterol* 2004[65]	54	No association
5-HT$_{3A}$ receptor Type A subunit (HT3A)	−42C>T	Kapeller et al, *Hum Mol Genet* 2008[66]	200	No overall IBS analysis performed; no association with IBS-C; C/T genotype increased frequency in UK IBS-D, not replicated in another cohort
	−25C>T	Kapeller et al, *Hum Mol Genet* 2008[66]	200	No overall IBS analysis performed; no association with IBS-D or IBS-C
	*70C>T	Kapeller et al, *Hum Mol Genet* 2008[66]	200	No overall IBS analysis performed; no association with IBS-D or IBS-C

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

Protein (gene)	Polymorphism	Reference	Case sample size	Finding summary
	*503C>T	Kapeller et al, *Hum Mol Genet* 2008[66]	200	No overall IBS analysis performed; no association with IBS-D or IBS-C
5-HT$_{3B}$ receptor Type B subunit	386A>C	Fukudo et al, *Neuroimage* 2009[25]	28	AC and CC subjects had differential brain activation compared with AA subjects
5-HT$_{3E}$ receptor Type E subunit (HT3E)	*76G>A	Kapeller et al, *Hum Mol Genet* 2008[66]	200	No overall IBS analysis performed; G/A genotype more common in IBS-D (odds ratio, 8.53; P<.05), replicated in separate cohort
5-HT$_{3C}$ receptor Type C subunit (5-HT3C)	489C>A	Kapeller et al, *Gastroenterology* 2009[67]	197	CC associated with female patients with IBS-D
	*115T>G	Kapeller et al, *Hum Mol Genet* 2008[66]	200	No overall IBS analysis performed; no association with IBS-D or IBS-C
	*138C>T	Kapeller et al, *Hum Mol Genet* 2008[66]	200	No overall IBS analysis performed; no association with IBS-D or IBS-C
	−189G>A	Kapeller et al, *Hum Mol Genet* 2008[66]	200	No overall IBS analysis performed; no association with IBS-D or IBS-C
Inflammatory markers				
Interleukin-10	−1082G>A	Gonsalkorale et al, *Gut* 2003[68]	230	AA or GA associated
		Van der Veek et al, *Gastroenterology* 2004[69]	111	No association
	−819C>T	Van der Veek et al, *Gastroenterology* 2004[69]	111	No association
Transforming growth factor β1	+869 T>C	Gonsalkorale et al, *Gut* 2003[68]	230	No association
	+915G>C	Gonsalkorale et al, *Gut* 2003[68]	230	No association
Tumor necrosis factor α	−308G>A	Van der Veek et al, *Gastroenterology* 2004[69]	111	GA associated
Adrenergic pathway				
Norepinephrine transporter (NET)	237G>C	Kim et al, *Gut* 2004[55]	276	Not detected
α_{2A}-adrenergic receptor (A2AR)	753C>G	Kim et al, *Gut* 2004[55]	276	Not detected
	1291C>G	Kim et al, *Gut* 2004[55]	276	Borderline association with IBS-C
		Sikander et al, *Clin Chim Acta* 2009[70]	151	Not associated with IBS overall; non-CC genotype associated with IBS-D
Alpha α_{2C}-adrenergic receptor (A2CR)	Del 332-335	Kim et al, *Gut* 2004[55]		
Other				
β3 subunit G protein (GNβ3)	825C>T	Holtmann et al, *Gastroenterology* 2004[71]	20	No association
		Andresen et al, *Gastroenterology* 2006[72]	233	No association
		Saito et al, *Neurogastroenterol Motil* 2007[59]	50	No association

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Protein (gene)	**Polymorphism**	**Reference**	**Case sample size**	**Finding summary**
		Saito et al, *Gastroenterology* 2009[64]	312	No association with IBS overall, but there was a gene-environment interaction between this SNP and prior gastrointestinal infection with IBS
Cholecystokinin type 1 receptor (*CCK1-R*)	266G>C 8641G>A 984T>C	Colucci et al, *Gastroenterology* 2009[73]	60	No association between these 3 SNPs with IBS or IBS symptom severity
Nav1.5 channel (*SCN5A*)	G298S	Saito et al, *Am J Physiol Gastrointest Liver Physiol* 2009[74]	49	Reported in 1 patient with IBS-D
43 genes (inflammatory, intestinal barrier, serotonin)	78 SNPs	Villani et al, *Gastroenterology* 2009[75]		SNPs in *CDH1*, *TLR9*, *and IL6* were associated with postinfectious IBS
43 genes (inflammatory, intestinal barrier, serotonin)	78 SNPs	Villani et al, *Gastroenterology* 2009[75]		SNPs in *IL10* and *HTR2A* associated with IBS-D in female patients; SNPs in *CDH1*, *TLR9, IL6* were not associated with non–postinfectious IBS

IBS-C, constipation-predominant irritable bowel syndrome.

Vaccine 29 (2011) 8988–8997

Contents lists available at SciVerse ScienceDirect

Vaccine

journal homepage: www.elsevier.com/locate/vaccine

Genetic polymorphisms in host antiviral genes: Associations with humoral and cellular immunity to measles vaccine

Iana H. Haralambieva[a,b], Inna G. Ovsyannikova[a,b], Benjamin J. Umlauf[a], Robert A. Vierkant[c], V. Shane Pankratz[c], Robert M. Jacobson[a,b,d], Gregory A. Poland[a,b,d,*]

[a] Mayo Clinic Vaccine Research Group, Mayo Clinic, Rochester, MN 55905, USA
[b] Program in Translational Immunovirology and Biodefense, Mayo Clinic, Rochester, MN 55905, USA
[c] Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, MN 55905, USA
[d] Department of Pediatric and Adolescent Medicine, Mayo Clinic, Rochester, MN 55905, USA

ARTICLE INFO

Article history:
Received 19 July 2011
Received in revised form 2 September 2011
Accepted 10 September 2011
Available online 20 September 2011

Keywords:
Single nucleotide polymorphisms
Haplotypes
Antiviral genes
Measles vaccine
Immunity

ABSTRACT

Host antiviral genes are important regulators of antiviral immunity and plausible genetic determinants of immune response heterogeneity after vaccination. We genotyped and analyzed 307 common candidate tagSNPs from 12 antiviral genes in a cohort of 745 schoolchildren immunized with two doses of measles–mumps–rubella (MMR) vaccine. Associations between SNPs/haplotypes and measles virus-specific immune outcomes were assessed using linear regression methodologies in Caucasians and African-Americans.

Genetic variants within the *DDX58*/RIG-I gene, including a coding polymorphism (rs3205166/Val800Val), were associated as single-SNPs ($p \leq 0.017$; although these SNPs did not remain significant after correction for false discovery rate/FDR) and in haplotype-level analysis, with measles-specific antibody variations in Caucasians (haplotype allele p-value = 0.021; haplotype global p-value = 0.076). Four *DDX58* polymorphisms, in high LD, demonstrated also associations (after correction for FDR) with variations in both measles-specific IFN-γ and IL-2 secretion in Caucasians ($p \leq 0.001$, $q = 0.193$). Two intronic *OAS1* polymorphisms, including the functional *OAS1* SNP rs10774671 ($p = 0.003$), demonstrated evidence of association with a significant allele-dose-related increase in neutralizing antibody levels in African-Americans. Genotype and haplotype-level associations demonstrated the role of *ADAR* genetic variants, including a non-synonymous SNP (rs2229857/Arg384Lys; $p = 0.01$), in regulating measles virus-specific IFN-γ Elispot responses in Caucasians (haplotype global p-value = 0.017). After correction for FDR, 15 single-SNP associations (11 SNPs in Caucasians and 4 SNPs in African-Americans) still remained significant at the q-value < 0.20.

In conclusion, our findings strongly point to genetic variants/genes, involved in antiviral sensing and antiviral control, as critical determinants, differentially modulating the adaptive immune responses to live attenuated measles vaccine in Caucasians and African-Americans.

© 2011 Elsevier Ltd. All rights reserved.

1. Introduction

The attenuated measles vaccine offered the opportunity to considerably reduce measles-related morbidity and mortality. However, measles is still a public health issue with approximately 164,000 deaths in 2008 worldwide [1] and with a relative increase in reported cases in several European countries and in the United States, including cases in individuals after two doses of MMR vaccine [2–6]. Therefore, establishing effective measles immunity across the population is crucial and requires providing both immunization coverage and surveillance to detect new cases and outbreaks [7]. Despite measles control (but not eradication), questions still remain regarding the documented 2–10% primary failure rate after vaccination, and the immunogenetic causes of vaccine response heterogeneity and/or vaccine failure. Deeper understanding of the observed inter-individual differences in immune response to measles vaccine is needed to provide new insights for improving vaccine-induced humoral, cellular, and innate immune responses, and for designing better vaccines [8,9].

The first line of defense, innate immunity, orchestrates host responses to prevent or reduce viral replication and spread until the adaptive immune system is switched on to eliminate the specific invading pathogen. Cellular viral sensors, type I and III interferons (IFNs, IFNα/β, IFN λ1, IFN λ2, IFN λ3) and IFN-triggered

---

[*] Corresponding author at: Mayo Clinic Vaccine Research Group, Mayo Clinic, Guggenheim 611C, 200 First Street SW, Rochester, MN 55905, USA.
Tel.: +1 507 284 4968; fax: +1 507 266 4716.
E-mail address: poland.gregory@mayo.edu (G.A. Poland).

0264-410X/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.vaccine.2011.09.043

antiviral proteins have long been recognized as crucial mediators of innate antiviral defense with important effects on the magnitude and quality of innate and adaptive immune responses. Important antiviral factors and pathways, such as the retinoic acid-inducible gene I protein (*RIG-I*/*DDX58*), melanoma differentiation associated protein-5 (*MDA-5*/*IFIH1*), mitochondrial antiviral signaling protein (*MAVS*/*VISA*/*IPS1*/*CARDIF*), dsRNA-activated protein kinase R (*PKR*/*EIF2AK2*), the 2′–5′-oligoadenylate-synthetase (*OAS*) – ribonuclease L (*RNASEL*) pathway, Mx protein GTPases, adenosine deaminase, RNA-specific 1(*ADAR*), the *ISG15* ubiquitin-like pathway and others play a role in viral sensing, control, pathogenesis and outcome of viral infections [10–13]. Polymorphisms in IFN-stimulated antiviral effector genes and cellular viral sensors are potential genetic determinants of immune response heterogeneity that may influence the immune responses to viral vaccines by altering the functionality and antiviral effects of the corresponding proteins. Genetic variants in these genes have been implicated as important regulators of immunity and host response to infection [14–24].

Bearing in mind the multifaceted interactions between viruses and factors of the innate immune system, our study sought to investigate for the first time the role of key antiviral effector proteins and cellular antiviral sensors as plausible contributors to immune response heterogeneity to live attenuated measles vaccine. For this reason, we performed a comprehensive candidate gene association study in a large racially diverse cohort of healthy schoolchildren after two doses of MMR vaccine. Our results suggest that multiple innate immunity genetic variants/genes are likely involved in modulating the adaptive immune responses to live attenuated measles vaccine in Caucasians and African-Americans.

2. Materials and methods

2.1. Study subjects and demographics

Our study cohort consisted of 2 independent age-stratified samples of randomly recruited healthy schoolchildren and young adults from Olmsted County, Minnesota, as described previously [25,26]. Briefly, cohort 1, enrolled in 2006–2007, comprised a sample of 440 subjects (age 11–19 years), from which 388 were eligible for the current study [14–16]. Cohort 2, enrolled in 2008–2009 year, consisted of 376 eligible subjects (age 11–22 years). All study participants resided in a community where no cases of measles infection had been reported during their lifetimes and all had received two doses of MMR-II vaccine (Merck, Whitehouse Station, N.J.), containing the Edmonston strain of measles virus ($TCID_{50} \geq 1000$). The Mayo Clinic Institutional Review Board granted approval for the study. Written, informed consent was obtained from parents/guardians as well as written assent from age-appropriate subjects at the time of enrollment in the study.

The demographic and clinical variables of the study cohort have been described previously [25,26]. Briefly, the study cohort ($n = 745$ subjects) used in our genetic association study, consisted of 417 males (55.97%), and 328 females (44.03%). The two most prevalent racial groups used were Caucasians (598, 80.27%) and African-Americans (89, 11.95%). Most of the study subjects were non-Hispanics (723, 97.05%). The median age at enrollment was 15 years (inter-quartile range/IQR 13; 17) and the median age at first and second immunization was 15 months (IQR 15; 16) and 5 years (IQR 4; 11), respectively. The median time from second immunization to enrollment and immunity measurements was 7.4 years (IQR 5.6; 9.2).

2.2. Immune measures

2.2.1. Plaque Reduction Microneutralization Assay (PRMN)

We measured neutralizing antibodies to measles virus using a high throughput fluorescence-based PRMN, as previously described [27], with the following modifications in the readout. We used an automated Olympus IX71 Fluorescent microscope system with Image-Pro Plus Software Version 6.3 (MediaCybernetics) to scan and count GFP-positive plaques [25]. Calculations of 50% end point titer (Neutralizing Dose, ND_{50}) were performed using Karber's formula and ND_{50} values were transformed into mIU/mL values using the 3rd WHO international anti-measles antibody standard (NIBSC code no. 97/648) as previously described [27].

2.2.2. Elispot

Human total IFN-γ Elispot kits (R&D Systems, Minneapolis, MN) were used to measure the number of IFNγ-producing cells, as previously described [28] following the manufacturer's protocol. Briefly, we stimulated subjects' peripheral blood mononuclear cells/PBMCs (or alternatively left them unstimulated), in triplicate, with the Edmonston strain of measles virus (multiplicity of infection, MOI = 0.5) and developed the reaction after 42 h incubation at 37 °C, in 5% CO_2. Phytohaemagglutinin/PHA (5 μg/mL) was used as a positive control. All plates were scanned and analyzed using the same counting parameters on an ImmunoSpot® S4 Pro Analyzer (Cellular Technology Ltd., Cleveland, OH, USA) using ImmunoSpot® version 4.0 software (Cellular Technology Ltd.).

2.2.3. ELISA

Secreted cytokines (Th1: IL-2, IFN-γ; Th2: IL-10; and innate/inflammatory: IL-6, TNF-α, IFN-α, IFN-λ1) were quantified in PBMC cultures after in vitro stimulation with live measles virus (stimulated and unstimulated in five replicate measurements), using pre-optimized conditions for MOI and incubation time for each cytokine, as previously described [29,30]. The different cytokines are measured using the following conditions: IFN-α and TNF-α MOI = 1.0, 24 h; IL-2 and IL-10 MOI = 0.5, 48 h; IL-6, IFN-γ and IFN-λ1 MOI = 1.0, 72 h. Secreted cytokine levels were measured with commercial kits according to manufacturers' recommendations (R&D Systems, Minneapolis, MN, for IFN-λ1; Mabtech, Cincinnati, OH, for IFNα, and BD Biosciences Pharmingen, San Diego, CA for the rest of the cytokines).

2.3. SNP selection and genotyping of SNPs in antiviral genes

The SNP selection approach and genotyping methodology have been published for our previous genetic association studies [14–16,26]. Briefly, we used the tagSNP selection approach based on linkage disequilibrium (LD) [31] to select SNPs from each of 12 antiviral genes, including the regions 5 kb upstream and downstream for each gene, from the source with the greatest number of SNPs from among the Hapmap Phase II (http://www.hapmap.org), Seattle SNPs (http://pga.mbt.washington.edu/) and NIEHS SNPs (http://egp.gs.washington.edu/). We binned SNPs with a minor allele frequency (MAF) ≥ 0.05, a pairwise LD threshold of $r^2 \geq 0.90$, and successful Illumina predictive genotyping scores, for the Caucasian sample and African-American sample separately, using the ldSelect program [31]. To accommodate a set of Illumina platform constraints, the SNP list was further post-processed and refined using the SNPPicker program. We also added a set of functional "obligate" SNPs (coding: nonsynonymous, synonymous, 5′ or 3′ untranslated regions) to the final list. A total of 348 SNPs across 12 antiviral genes were selected based on this approach. The nomenclature used for the description of the polymorphisms follows that described by den Dunnen and Antonarakis [32].

We genotyped the selected 348 SNPs using two custom Illumina GoldenGate SNP panels (Illumina Inc., San Diego, CA) for 1536 and 768 SNPs, which were used for genotyping all candidate SNPs from a large population-based measles vaccine study. All DNA samples ($n = 764$) were genotyped following the manufacturer's protocol and as previously described [14].

Seven hundred and sixty four ($n = 764$) subjects were initially genotyped for 348 known SNPs in 12 antiviral genes with a documented role in antiviral immunity (*DDX58/RIG-I*, *MAVS/VISA*, *IFIH1/MDA5*, *MX1*, *MX2*, *OAS1*, *OAS2*, *OAS3*, *PKR/EIF2AK2*, *RNASEL*, *ADAR* and *ISG15*). The genotyping locus success rate was 94.55%, sample success rate was 98.75% (including replicate samples), and reproducibility was 100%. Due to insufficient/inadequte DNA quality, genotyping failure or low call rates, nineteen DNA samples failed, leaving 745 subjects for further analysis. Of the 348 SNPs included for genotyping, eighteen SNPs failed the initial quality control because of failure to amplify, poor clustering or multiple replicate errors. 330 SNPs (95.12%) yielded genotyping data. Additional SNPs were also excluded from analysis based on low MAF < 0.05 ($n = 20$) or based on being monomorphic ($n = 3$). A total of 307 SNPs were used for the final race-specific analysis in 598 Caucasians and 89 African-Americans.

2.4. Statistical methods

The statistical methods described herein are similar to those published for our previous genetic association studies [14–16,26]. Data were descriptively summarized using frequencies and percentages for categorical variables, and medians and inter-quartile ranges for continuous variables. Assessments of cytokine secretion and cell-mediated immunity (CMI) measures resulted in multiple recorded values per outcome both prior to and after stimulation with measles virus. For descriptive purposes, a single response measurement per individual was obtained for each of these outcomes by subtracting the median of the unstimulated values from the median of the stimulated values. Assessments of antibody levels resulted in only one recorded value per individual.

Observed genotypes were used to estimate allele frequencies for each SNP and departures from Hardy-Weinberg equilibrium (HWE) were assessed using either a Pearson goodness-of-fit test or, for SNPs with a minor allele frequency (MAF) of less than 5%, a Fisher exact test [33]. Estimates of pair-wise linkage disequilibrium (LD) based on the r-squared statistic were obtained using Haploview software, version 3.32 [34].

SNP associations with immune response outcomes were evaluated using regression models. Simple linear regression was used for measles antibody levels, for which we had only one measured value per individual. Repeated measures approaches were implemented for the cytokine secretion and IFN-γ Elispot variables, simultaneously modeling the multiple observed measurements and using an unstructured within-person variance–covariance matrix to account for within-subject correlations. This was achieved by including the genotype variable in the regression model, together with a variable representing stimulation status. The resulting genotype-by-stimulation status interaction was then tested for statistical significance. Tests of association assumed an ordinal (log-additive) SNP effect using simple tests for trend. Because of the lack of seronegative subjects (PRMN titer < 1:8) in our cohort and the low number of subjects (68) with antibody concentrations less than the protective threshold of 210 mIU/mL (corresponding to PRMN titer of 120) [25], our statistical approach rationale was to assess associations between SNPs and quantitative humoral/cellular immune outcomes across the entire cohort using linear regression methodologies rather than assess the distribution of SNPs across extremes of humoral/cellular immune responses (such as seronegative subjects and/or subjects with antibody concentrations less than the protective titer).

To further explore genomic regions containing statistically significant single-SNP effects, we performed post hoc haplotype analyses. Posterior probabilities of all possible haplotypes for an individual, conditional on the observed genotypes, were estimated using an expectation–maximization (EM) algorithm [35]. This information was used to define haplotype design variables that estimated the number of each of the haplotypes carried by an individual. Analyses were performed on all resulting common haplotypes (those with an estimated frequency of >1%) using simple least squares linear regression approach for antibody levels and repeated measures approaches for the CMI variables. Differences in immune response among common haplotypes were first assessed globally and simultaneously tested for statistical significance using a multiple degree-of-freedom test. Following these global tests, we examined individual haplotype effects. Each haplotype was included in a separate regression analysis, effectively comparing immune response levels for the haplotype of interest against all others combined. Due to phase ambiguity, haplotype-specific medians and inter-quartile ranges could not be calculated. Thus, descriptive summaries were represented using the corresponding t-statistics.

All of the association analyses described above were separately performed within the two most prevalent racial groups (Caucasians and African-Americans). Each analysis also adjusted for gender, age at first measles vaccination, age at second measles vaccination, age at enrollment (thus accounting for the time since last vaccination to enrollment/blood draw and immunity measurements), and cohort status (cohort 1 vs. cohort 2). We used an inverse normal transformation for all cytokine secretion and IFN-γ Elispot outcome variables, and a log transformation for the antibody response measure, in all formal tests to account for the skewed nature of the data. Because of the large number of statistical tests, we computed q-values for each SNP to estimate the corresponding false discovery rates (FDR) as per Storey and Tibshirani [36,37]. We considered SNPs as being significant if their q-value was less than 0.2 [38]. However, we also reported single-SNP associations as being of interest for further follow-up if the q-value was less than 0.5. This corresponded to single-SNP p-values that were roughly ≤ 0.01. All statistical tests were two-sided and, unless otherwise indicated, all analyses were carried out using the SAS software system (SAS Institute, Inc., Cary, N.C.).

3. Results

3.1. Immune responses for the study cohort

The immune characteristics of the study cohort are summarized in Table 1. The median level of measles-specific neutralizing antibodies (humoral immunity) for the study cohort was 846 mIU/mL (inter-quartile range/IQR: 418, 1772) and the median IFN-γ Elispot response (cellular immunity) was 36 IFN-γ-positive spots per 2×10^5 cells (IQR: 12, 69). The median measles-specific IFN-γ secretion level for the cohort was 67 pg/mL (IQR: 35, 120), the median measles-specific IL-2 secretion level was 38 pg/mL (IQR: 21, 64), and the median measles-specific IL-10 secretion level was 18 pg/mL (IQR: 11, 28) [26].

3.2. Genetic associations

3.2.1. Associations between SNPs in antiviral genes and humoral immune responses after measles vaccination

Race-specific analysis in the Caucasian group identified two significant associations between SNPs located in antiviral genes and neutralizing antibody levels after measles vaccination ($p \leq 0.01$), as well as one suggestive SNP association with $p = 0.017$ (Table 2). An intronic SNP in the *DDX58* (RIG-I) gene (rs11795343) was associated with an allele dose-related increase (19%, calculated as the ratio, minus one, of the immune outcome observed for the homozygous minor allele genotype to the immune outcome observed for the homozygous major allele genotype, multiplied

Table 1
Immune characteristics of the study cohort.

Response type	Cytokine	No. of subjects	Median response (IQR)[a]	No. subjects positive (%)[b]
Neutralizing antibody	–	744	846 (418; 1 772)	686 (92.2)
Th1	IL-2	739	38 (21; 64)	724 (98.0)
	IFNγ	737	67 (35; 120)	716 (97.2)
Th2	IL-10	740	18 (11; 28)	726 (98.1)
Innate/inflammatory	IL-6	737	355 (248; 461)	726 (98.5)
	TNFα	732	14 (9; 19)	722 (98.6)
	IFNα	734	551 (273; 1 025)	729 (99.3)
	IFNλ1	738	34 (14; 73)	666 (90.2)
Elispot	IFNγ	707	36 (12; 69)	676 (95.6)

[a] Humoral response is defined as the measles-specific mIU/mL titer in PRMN assay. Cytokine response is defined as the subject-specific median measles-stimulated response (measured in five replicates) minus the median unstimulated response (also measured in five replicates). Elispot response is defined as the subject-specific median measles-stimulated response (measured in triplicate) minus the median unstimulated response (also measured in triplicate). Values reported are in mIU/mL for antibody responses, pg/mL for secreted cytokines, and cytokine-positive spots per 2×10^5 cells for Elispot responses ± IQR, inter-quartile range.

[b] For antibody responses subjects reported as positive are with positive protective titers higher than the protective threshold of 210 mIU/mL (corresponding to ND_{50} 120). Subjects are considered to have a positive cytokine response if the median of the stimulated cells is larger than the median of the unstimulated cells.

by 100%) in antibody levels ($p = 0.01$). This intronic SNP was found to be in LD ($r^2 = 0.64$) with a synonymous *DDX58* SNP (rs3205166/Val800Val), which also demonstrated an allele dose relationship with higher neutralizing antibody levels (approximately 40% increase in antibodies), although this association did not reach statistical significance at $p \leq 0.01$ ($p = 0.017$, Table 2). In addition, eight more *DDX58* SNPs displayed suggestive associations with antibody levels in this racial group ($p < 0.05$, data not shown). An intronic genetic variant within the *VISA* gene demonstrated an association with antibody levels in the Caucasian group ($p = 0.003$).

The analysis restricted to the African-American subjects revealed three significant intronic SNP associations with measles-specific humoral immune response ($p \leq 0.01$, Table 2). Importantly, this analysis identified an association for a functional genetic variant rs10774671 located in intron 5 of *OAS1*, at the splice-acceptor site [39]. The minor alleles of the two *OAS1* intronic SNPs, rs10774671 and rs10744785, were associated with a significant increase (more than a 3-fold increase for rs10774671 and 4.9-fold increase for rs10744785) in measles neutralizing antibodies in an allele dose-related manner ($p \leq 0.003$, Table 2).

However, after adjusting for false discovery rate (FDR), none of the identified SNPs listed in Table 2 remained significant at the q-value = 0.20 [38], although all reported SNP associations demonstrated q-values < 0.5 and two SNP associations demonstrated q-values < 0.25.

3.2.2. Associations between SNPs in antiviral genes and cellular immune responses (INF-γ Elispot) after measles vaccination

Race-specific analysis in the Caucasian group revealed six significant associations between genetic variants in two antiviral genes (*OAS3* and *ADAR*) and measles-specific IFN-γ Elispot responses ($p \leq 0.01$, Table 3). Four *ADAR* SNPs in high LD ($r^2 \geq 0.96$), including a non-synonymous *ADAR* SNP (rs2229857/Lys384Arg), demonstrated an allele dose-related increase (29% for rs2229857) in IFN-γ Elispot responses in this racial group with the representation of the minor allele ($p \leq 0.01$, Table 3). Furthermore, we observed three additional *ADAR* SNPs, which demonstrated suggestive associations

Table 2
SNPs in antiviral genes associated with measles virus-specific neutralizing antibody responses.

Gene	SNP ID	Location/function	Geno type[a]	*N* (%)[a]	Median Ab level mIU/mL (IQR)[b]	*p*-Value[c]	*q*-Value[d]
Analysis in Caucasians							
VISA	rs6037678	Intron	CC	558 (93)	804 (409; 1 718)	0.003	0.23
			CG	40 (7)	1 419 (880; 2 144)		
			GG	0 (0)	(–)		
DDX58	**rs11795343**	Intron	AA	203 (34)	**759** (389; 1 528)	0.01	0.408
			AG	305 (51)	**898** (455; 1 806)		
			GG	89 (15)	**904** (497; 1 968)		
DDX58	**rs3205166**	Coding	AA	234 (39)	**797** (410; 1 726)	0.017	0.438
		Val800Val	AC	276 (46)	**839** (410; 1 711)		
			CC	87 (15)	**1 109** (551; 2 096)		
Analysis in African-Americans							
OAS1	**rs10744785**	Intron	GG	24 (27)	**398** (194; 637)	0.001	0.248
			GC	46 (52)	**857** (329; 2 042)		
			CC	19 (21)	**1 967** (918; 3 564)		
OAS1	**rs10774671**	Intron	GG	36 (40)	**492** (221; 1 138)	0.003	0.403
			GA	38 (43)	**857** (410; 2 228)		
			AA	15 (17)	**1 690** (918; 2 981)		
MX2	rs2838034	Intron	GG	60 (67)	681 (265; 1 948)	0.01	0.470
			GA	24 (27)	565 (283; 2 030)		
			AA	5 (6)	2 811 (2 042; 4 172)		

Bolded SNPs are associated with an allele dose-related increase (with the representation of the minor allele) in antibody levels.

[a] Values are presented as homozygous major allele/heterozygous/homozygous minor allele. *N* indicates the number of subjects with the specific genotype, % indicates genotype counts in percentages.

[b] Results are presented as median neutralizing antibody (Ab) levels in mIU/mL ± IQR (inter-quartile range), as measured by the fluorescence-based measles virus-specific PRMN assay.

[c] Ordinal *p*-value from the linear regression analysis adjusting for age at enrollment, gender, age at first and second immunization, and cohort status (cohort 1 vs. cohort 2).

[d] Corresponding *q*-values, adjusting for FDR.

Table 3
SNPs in antiviral genes associated with measles virus-specific IFN-γ Elispot responses.

Gene	SNP ID	Location/function	Geno type[a]	N (%)[a]	Median IFNγ Elispot (IQR)[b]	p-Value[c]	q-Value[d]
Analysis in Caucasians							
OAS3	**rs2072133**	3′UTR	AA	373 (66)	**36** (16; 70)	0.007	0.328
			AG	170 (30)	**37** (11; 71)		
			GG	19 (3)	**40** (10; 85)		
ADAR	rs1127317	3′UTR	AA	294 (52)	36 (12; 67)	0.009	0.387
			AC	217 (39)	36 (16; 73)		
			CC	52 (9)	45 (24; 89)		
ADAR	rs1127326	3′UTR	GG	293 (52)	36 (12; 67)	0.01	0.4
			GA	217 (39)	36 (16; 73)		
			AA	53 (9)	45 (22; 87)		
OAS3	rs12818640	3′ intergenic	AA	370 (66)	36 (16; 70)	0.01	0.408
			AG	170 (30)	36 (11; 71)		
			GG	18 (3)	33 (10; 85)		
ADAR	**rs2229857**	Coding	GG	291 (52)	**35** (11; 68)	0.01	0.408
		Lys384Arg	GA	219 (39)	**36** (16; 73)		
			AA	52 (9)	**45** (21; 87)		
ADAR	**rs3738029**	Intron	GG	292 (52)	**35** (11; 68)	0.01	0.408
			GA	219 (39)	**37** (16; 73)		
			AA	52 (9)	**45** (21; 87)		
Analysis in African-Americans							
EIF2AK2	**rs12712526**	Intron	AA	23 (26)	**30** (12; 61)	0.001	0.248
			AG	43 (49)	**29** (5; 47)		
			GG	21 (24)	**10** (2; 39)		
DDX58	rs7022323	Intron	AA	32 (37)	12 (3; 29)	0.004	0.447
			AG	40 (46)	35 (5; 71)		
			GG	15 (17)	33 (3; 49)		
MX2	**rs466092**	Intron	GG	39 (45)	**11** (2; 43)	0.006	0.447
			GA	35 (40)	**27** (10; 51)		
			AA	13 (15)	**30** (12; 43)		
OAS3	rs2072136	Coding	GG	71 (82)	26 (4; 49)	0.007	0.447
		Ser567Ser	GA	15 (17)	10 (2; 47)		
			AA	1 (1)	29 (29; 29)		
OAS2	rs11513733	Intron	AA	84 (97)	27 (4; 48)	0.009	0.454
			AT	3 (3)	9 (0; 11)		
			TT	0 (0)	(–)		

Bolded SNPs are associated with an allele dose-related increase/decrease in median IFN-γ Elispot responses.

[a] Values are presented as homozygous major allele/heterozygous/homozygous minor allele. *N* indicates the number of subjects with the specific genotype, % indicates genotype counts in percentages.

[b] Results are presented as median IFNγ-positive spots per 2×10^5 cells ± IQR (inter-quartile range), as measured by the total IFNγ Elispot assay.

[c] Ordinal *p*-value from the repeated measures linear regression analysis adjusting for age at enrollment, gender, age at first and second immunization, and cohort status.

[d] Corresponding *q*-values, adjusting for FDR.

with the immune outcome in the Caucasian group ($p < 0.05$, data not shown).

The analysis in the African-American group revealed five associations for SNPs belonging to five antiviral genes with measles-specific cellular immunity ($p \leq 0.009$, Table 3). A synonymous *OAS3* polymorphism (rs2072136/Ser567Ser) was found to be associated with IFN-γ Elispot responses in this racial group ($p = 0.007$, Table 3). The minor alleles of two intronic SNPs, *EIF2AK2* SNP rs12712526 ($p = 0.001$) and *MX2* SNP rs466092 ($p = 0.006$), demonstrated significant allele dose-related variations (decrease and increase, respectively) in the immune outcome (Table 3).

After adjusting for false discovery rate (FDR), none of the identified SNPs in Table 3 remained significant at the *q*-value = 0.20 [38], although all reported SNP associations demonstrated *q*-values < 0.5 and one SNP association demonstrated *q*-values < 0.25.

3.2.3. Associations between DDX58 and ADAR haplotypes and measles vaccine-induced immunity in Caucasians

To further explore genomic regions enriched with statistically significant ($p \leq 0.01$) and suggestive ($p < 0.05$, data not shown) single-SNP associations, we performed haplotype analyses for the *DDX58* and *ADAR* gene regions.

The haplotype level analysis revealed that genetic variants/haplotypes within the *DDX58* and *ADAR* genes (Haploview output presented in Fig. 1) were associated with measles-specific humoral and cellular immunity (IFN-γ Elispot), respectively, in Caucasians, but not in African-Americans (*DDX58* haplotype allele *p*-value = 0.021; *ADAR* global *p*-value = 0.017, Table 4). Particularly, the *DDX58* haplotype GGAAAAAAAA was associated with decreased antibody levels (haplotype *t*-statistic = −2.31; allele *p*-value = 0.021) compared to all other *DDX58* haplotypes, while the *ADAR* haplotype ATAATCGGGAAAGGAA was associated with increased IFN-γ Elispot response (haplotype *t*-statistic = 2.82; allele *p*-value = 0.005) in the Caucasian group (Table 4).

3.2.4. Associations between SNPs in antiviral genes and secreted cytokines after measles vaccination

We found 58 and 47 significant associations ($p \leq 0.01$) in the Caucasian and African-American group, respectively, between genetic variants in antiviral effector genes and measles-specific secretion of Th1, Th2 and innate/inflammatory cytokines, however most of the SNP associations (with cytokine secretion immune outcomes) did not meet the FDR threshold of 0.20 [38]. Analysis in the Caucasian group revealed seven significant SNP associations with measles-specific IFN-γ secretion, which remained significant at

Table 4
DDX58 and *ADAR* haplotype associations with measles-specific immune responses in Caucasians.

Haplotype	Frequency	Test statistic (haplotype *t*-statistic)[c]	Allele *p*-value[d]	Global *p*-value[d]
DDX58 haplotype associations with antibody levels				
ACAGGGAAAA[a]	0.295	−1.21	0.229	
ACAGGGAAAC[a]	0.016	0.41	0.680	
GGAAAAAAAA[a]	0.208	−2.31	**0.021**	0.076
GGAGGGAGGA[a]	0.092	0.51	0.609	
GGAGGGGAAC[a]	0.038	1.04	0.297	
GGGGGGGGGC[a]	0.305	1.59	0.113	
ADAR haplotype associations with IFNγ Elispot responses				
ATAATCGGGAAAGGAA[b]	0.072	2.82	**0.005**	
GAAGAAAAGGGCAAAA[b]	0.031	1.00	0.320	
GAAGAAAAGGGCAAAG[b]	0.247	−0.97	0.331	
GAAGTAGAGGGAGGCA[b]	0.183	−1.43	0.153	**0.017**
GAGGTAAAGGGCAAAG[b]	0.210	0.04	0.971	
GTAATCGGAAAAGGAA[b]	0.203	1.36	0.173	
GTGGTAAGGGGAGGAA[b]	0.013	−1.62	0.106	

[a] *DDX*58 genetic variants from left to right: rs7042042, rs10970985, rs10970986, rs12006123, rs12555727, rs17289116, rs7865082, rs10738889, rs10738890, and rs3205166 (all genotyped in the current study).

[b] *ADAR* genetic variants from left to right: rs3811450, rs6656743, rs9427092, rs1127326, rs9616, rs1127317, rs1127311, rs2131902, rs3738032, rs2229857, rs3738029, rs3766922, rs12125166, rs9427100, rs2335230, and rs884618 (all genotyped in the current study).

[c] Haplotype effects (for haplotypes with estimated frequencies ≥0.01) are calculated using the haplotype *t*-statistic, which shows the direction and magnitude of the estimated haplotypic effect on the immune measure.

[d] Allele *p*-values compare individual haplotypes to all other haplotypes combined; global *p*-values reflect differences in immune response among common haplotypes as assessed globally and tested for statistical significance using a multiple degree-of-freedom test. Statistically significant *p*-values for haplotype-level associations ($p < 0.05$) are highlighted in bold. Analyses are adjusted for age at enrollment, gender, age at first and second MMR immunization and cohort status (cohort 1 or cohort 2).

the *q*-value = 0.20 [38] after adjusting for FDR ($p \leq 0.001$, $q = 0.193$; Table 5). Two *MX2* intronic SNPs demonstrated an allele–dose relationship with lower IFN-γ secretion levels ($p \leq 0.001$). In addition, the minor alleles of four *DDX58* SNPs (in high LD; for rs12555727, 12006123 and rs17289116, $r^2 = 1$) were associated with a significant allele dose-related decrease (approximately 30%) in secreted IFN-γ in the Caucasian group (Table 5). Interestingly, the same four *DDX58* SNPs demonstrated a more than 2-fold reduction in measles-specific IL-2 secretion levels (with the representation of the minor allele) in the Caucasian group and still remained significant after adjusting for FDR ($p \leq 0.001$, $q = 0.193$, Table 6). Of note, three of the *DDX58* SNPs (rs12555727, 12006123 and rs17289116, $r^2 = 1$) were also found to be associated at the haplotype level with variations in measles-specific humoral immunity in Caucasians (Table 4).

The analysis in the African-American group revealed four significant SNP associations with measles-specific IL-10 secretion, which remained significant at the *q*-value = 0.20 [38] after adjusting for FDR ($p \leq 0.0003$, $q \leq 0.192$; Supplementary Table 1). In particular, the minor alleles of two SNPs in the 3′ intergenic region of the *MX1*

Table 5
SNPs in antiviral genes associated with measles virus-specific IFN-γ secretion in Caucasians.

Gene	SNP ID	Location/function	Geno type[a]	*N* (%)[a]	Median IFN-γ pg/mL (IQR)[b]	*p*-Value[c]	*q*-Value[d]
*MX*2	**rs442014**	Intron	AA	456 (77)	**71** (38; 126)		
			AG	131 (22)	**67** (29; 109)	0.0004	0.193
			GG	5 (1)	**14** (11; 38)		
*DDX*58	**rs12555727**	3′UTR	GG	363 (62)	**71** (37; 139)		
			GA	199 (34)	**70** (36; 107)	0.001	0.193
			AA	28 (5)	**50** (26; 96)		
*DDX*58	**rs12006123**	3′UTR	GG	364 (62)	**71** (37; 139)		
			GA	199 (34)	**70** (36; 107)	0.001	0.193
			AA	28 (5)	**50** (26; 96)		
*MX*2	**rs441437**	Intron	AA	432 (73)	**72** (38; 126)		
			AG	153 (26)	**61** (30; 109)	0.001	0.193
			GG	7 (1)	**38** (11; 161)		
*DDX*58	**rs17288914**	3′ Intergenic	AA	364 (61)	**71** (37; 139)		
			AG	199 (34)	**70** (36; 107)	0.001	0.193
			GG	29 (5)	**54** (27; 95)		
*DDX*58	**rs17289116**	Intron	GG	364 (61)	**71** (37; 139)		
			GA	199 (34)	**70** (36; 107)	0.001	0.193
			AA	29 (5)	**54** (27; 95)		
*OAS*1	rs4766662	Intron	CC	331 (56)	72 (39; 126)		
			CA	229 (39)	61 (32; 116)	0.001	0.193
			AA	31 (5)	88 (35; 144)		

Bolded SNPs are associated with an allele dose-related decrease in IFN-γ secretion levels. Only polymorphisms with $p \leq 0.01$ and $q < 0.2$ are presented in the table.

[a] Values are presented as homozygous major allele/heterozygous/homozygous minor allele. *N* indicates the number of subjects with the specific genotype, % indicates genotype counts in percentages.

[b] IQR, inter-quartile range, values are median levels of the cytokine in pg/mL as measured by ELISA.

[c] Ordinal *p*-value from the repeated measures linear regression analysis adjusting for age at enrollment, gender, age at first and second MMR immunization, and cohort status (cohort 1 or cohort 2).

[d] Corresponding *q*-values, adjusting for FDR.

Table 6
SNPs in antiviral genes associated with measles virus-specific IL-2 secretion in Caucasians.

Gene	SNP ID	Location/Function	Geno type[a]	N (%)[a]	Median IL-2 pg/mL (IQR)[b]	p-Value[c]	q-Value[d]
*DDX*58	**rs12555727**	3′UTR	GG	364 (61)	**44** (25; 71)		
			GA	200 (34)	**37** (20; 63)	0.0003	0.193
			AA	28 (5)	**21** (12; 42)		
*DDX*58	**rs12006123**	3′UTR	GG	365 (61)	**44** (25; 70)		
			GA	200 (34)	**37** (20; 63)	0.0003	0.193
			AA	28 (5)	**21** (12; 42)		
*DDX*58	**rs17288914**	3′ Intergenic	AA	365 (61)	**44** (25; 70)		
			AG	200 (34)	**37** (20; 63)	0.001	0.193
			GG	29 (5)	**21** (14; 43)		
*DDX*58	**rs17289116**	Intron	GG	365 (61)	**44** (25; 70)		
			GA	200 (34)	**37** (20; 63)	0.001	0.193
			AA	29 (5)	**21** (14; 43)		

Bolded SNPs are associated with an allele dose-related decrease in IL-2 secretion levels. Only polymorphisms with $p \leq 0.01$ and $q < 0.2$ are presented in the table.

[a] Values are presented as homozygous major allele/heterozygous/homozygous minor allele. *N* indicates the number of subjects with the specific genotype, % indicates genotype counts in percentages.

[b] IQR, inter-quartile range, values are median levels of the cytokine in pg/mL as measured by ELISA.

[c] Ordinal *p*-value from the repeated measures linear regression analysis adjusting for age at enrollment, gender, age at first and second MMR immunization, and cohort status (cohort 1 or cohort 2).

[d] Corresponding *q*-values, adjusting for FDR.

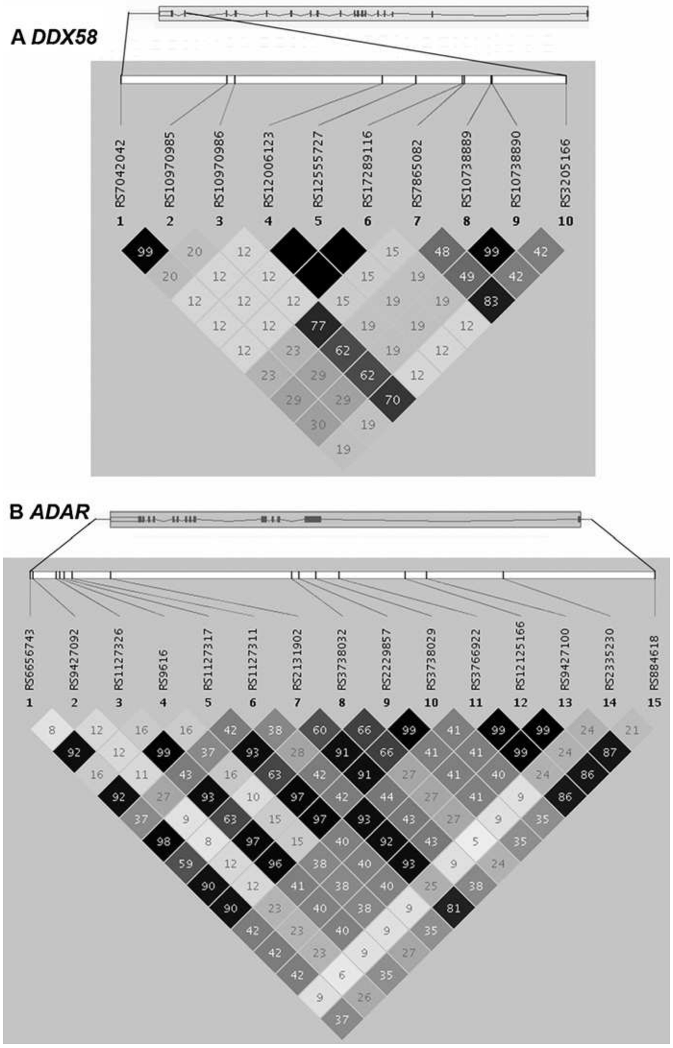


Fig. 1. Haplotype block structure of the *DDX58* and *ADAR* genetic variants in the study cohort (Caucasian sample). The schematic gene representation and LD block structure of (A) *DDX58* and (B) *ADAR* genetic variants is depicted. The LD block structure was analyzed using Haploview software, version 3.32. All SNPs presented in A and B were directly genotyped in the study. The r^2 color scheme is: white ($r^2 = 0$), shades of grey ($0 < r^2 < 1$), black ($r^2 = 1$). The numbers report the r^2 value multiplied by 100.

gene (in high LD) were associated with a significant increase in IL-10 secretion in an allele dose-related manner ($p = 0.0002$, $q = 0131$; Supplementary Table 1).

To summarize, we identified two *DDX58* and four *ADAR* genetic variants, as well as *DDX58* and *ADAR* haplotypes, associated with neutralizing antibody levels and measles-specific cellular immunity (IFN-γ Elispot), respectively, in our Caucasian group. Four *DDX58* polymorphisms (associated with neutralizing antibody variations at the haplotype level) demonstrated significant associations with variations in both IFN-γ and IL-2 secretion in Caucasians. Two *OAS1* genetic variants, including the functional *OAS1* SNP rs10774671, were associated with humoral immune response variations in our African-American group.

4. Discussion

Various genetic variants and genes have been implicated as important regulators of immunity to viral vaccines (including measles vaccine), such as cytokine and cytokine receptors, toll-like receptors (TLRs) and signaling molecules, viral receptors (such as CD46, SLAM and DC-SIGN for measles virus) and vitamin A and D receptors [15,16,26,38,40–47]. The objective of our population-based study was to comprehensively evaluate the influence of human genetic variation in key host antiviral sensor and antiviral effector genes on immune response heterogeneity after measles vaccination.

Genetic polymorphisms (SNPs) in genes related to innate immunity, such as the interferon β gene (*IFNB1*), type I IFN receptor genes *IFNAR1* and *IFNAR2*, the antiretroviral gene *TRIM5* and *TLRs* are known to be associated with variations in immune responses to measles and rubella vaccines [15,16,38,43,48]. We have previously demonstrated that specific variations in antiviral sensor genes (*DDX58*/*RIG-I* and *MAVS*/*VISA*) and key IFN-induced antiviral genes (particularly polymorphisms in the *OAS* gene cluster) were associated with differences in rubella vaccine-induced adaptive immunity [14–16], while little is known regarding the importance of these genes or variations in measles-specific immunity. Furthermore, multiple previous genetic association studies revealed that polymorphisms in the *MX1*, *OAS1*, *OASL* and *PKR* genes correlate with response to IFN therapy and/or susceptibility to viruses such as WNV, HCV, HBV, SARS-coV, and measles virus [17–24].

Our genetic data from single SNP and haplotype analyses point to the possible involvement of genetic variants within the *DDX58* (RIG-I) gene, including a coding polymorphism (rs3205166/Val800Val), in antibody variations following measles vaccine in Caucasians (haplotype allele p-value = 0.021). Mitochondrial antiviral pathways involve the detection of viral RNA by cytosolic helicases belonging to the RIG-I-like helicase family (RIG-I/*DDX58* and MDA5/*IFIH1*), and downstream signaling through interaction with the IPS-1 (*VISA*/*MAVS*/*CARDIF*) adaptor protein [49,50]. In addition, vitamin A and its metabolites, which are known to affect measles replication and antibody responses after vitamin supplementation, are linked with RIG-I/DDX58 through RARα and IFN type I-dependent mechanism [51,52]. Genetic variants within these antiviral sensor genes were associated with different measles-specific immune outcomes in our study (including common *DDX58* genetic variants regulating both antibody responses, and measles-specific IFN-γ and IL-2 secretion in the Caucasian group), which is consistent with a recent report of the recognition of measles virus by RIG-I and MDA5 to instigate interferon-specific innate immune responses [53]. Interestingly, a common coding functional *DDX58* polymorphism (rs10813831/Arg7Cys), known to modulate *IFNB1* and *DDX58* transcription in dendritic cells [54], and suggested as a plausible regulator of rubella vaccine-induced humoral immunity [15], was associated with measles-specific IFN-α secretion in the African-American sample, although did not remain significant after adjusting for FDR ($p = 0.005$, $q = 0.447$, data not shown).

In addition, our analysis of genetic associations with measles-specific vaccine responses revealed enriched representation of genetic variants within the OAS cluster of genes: *OAS1*, *OAS2* and *OAS3* (Tables 2, 3 and 5, Supplementary Table 1). The interferon-induced OAS1 protein is known to catalyze the synthesis of 2′,5′-oligoadenylates, which activate RNase L, resulting in antiviral (viral RNA degradation, inhibition of viral replication) and pro-apoptotic activities, and genetic polymorphisms in the coding genes have been linked to various immunological and infectious outcomes [21–24]. The involvement of *OAS* genetic variants in regulating measles vaccine–immune responses is highly plausible bearing in mind that *OAS* genetic variations are likely involved in modulating the adaptive immune responses to live attenuated rubella vaccine [14]. Particularly convincing is the association of *OAS1* SNP rs10774671, a polymorphism with established functionality, with variations in measles-specific humoral immune responses (more than 3-fold increase associated with the homozygous minor allele genotype) in the African-American sample ($p = 0.003$). This genetic variant is located in intron 5, at the splice-acceptor site, and has been demonstrated to regulate splicing, resulting in isoforms with different enzyme activity [39,55]. Recent data from the literature suggests that genetic variation in *OAS1* (rs10774671) can alter viral replication in human lymphoid tissue and is a risk factor for infection with West Nile virus [39]. In addition, our previous work links this same polymorphism in addition to other *OAS* polymorphisms with variations in rubella vaccine-induced cytokine secretion [14].

Another interesting antiviral gene potentially associated with measles virus-specific immune variations is *ADAR*, which is involved in RNA editing (that substitutes adenosines by inosines) of cellular mRNA and viral dsRNA targets. Our data show evidence for an association of a nonsynonymous SNP (rs2229857/Arg384Lys, in LD with other significant *ADAR* SNPs, $r^2 \geq 0.96$) with measles virus-specific cellular immunity (IFN-γ Elispot responses) in Caucasians ($p = 0.01$, Table 3). Consistent with the current findings, our previous rubella vaccine study demonstrated an increase in rubella-specific IFN-γ responses ($p = 0.021$) with the representation of the minor allele genotype for this *ADAR* polymorphism [14], which further suggests the involvement of this gene/genetic variant in immune response variations after vaccination. In addition, other regulatory and intronic SNPs within the same gene were associated at the genotype and haplotype level with variations in measles-specific cellular immunity, particularly in the Caucasian group (haplotype global p-value = 0.017, Table 4), implying cooperative effects of multiple SNPs in the genetic regulation of the immune outcome. This is in concert with the identification of ADAR as a critical determinant of local and systemic inflammatory responses through RNA editing and modulation of protein production [56]. It is important, however, to note that we did not find similarities between the genes/genetic variants associated with measles-specific IFN-γ Elispot responses and IFN-γ cytokine secretion in our analyses in Caucasians, and while the genetic associations with secreted IFN-γ remained significant after adjusting for FDR, none of the SNPs associated with IFN-γ Elispot responses remained significant after adjustment. Both our IFN-γ assays quantify IFN-γ responses, but as reported previously these assays do not correlate well with each other [57]. IFN-γ Elispot measures the number of cytokine-producing cells, while the IFN-γ ELISA measures the cytokine protein concentration released in the culture supernatant by virus-stimulated cells. We speculate that the lack of correlation (or weak correlation) between the two assays [57] is likely the main reason for the different genetic associations found in our analyses for these two immune outcomes.

Our findings point to the plausible role of several interrogated innate immunity genes in the observed inter-individual variations following measles vaccination, although some of the observations should be interpreted with caution since many of the single-SNP associations did not remain significant after adjusting for FDR at $q < 0.20$ [38].

The identified genes influencing vaccine-induced measles immunity are all interconnected components of the innate antiviral responses targeting viral nucleic acids. Viral nucleic acids stimulate the OAS-RNase L pathway resulting in RNA degradation. Viral RNA is also targeted by ADAR, which deaminates viral replication intermediates. In addition, viral RNA activates RNA helicase-like receptors, including RIG-I, to initiate signaling and transcription of IFNs (through IRFs), which feed back through their receptors to increase the expression of antiviral proteins, thus augmenting the antiviral response [10,58]. Hence, our findings imply that key genetic variants/genes involved in the initial antiviral response to viral nucleic acids are modulating immune responses after vaccination with a live viral vaccine, likely due to alterations in viral sensing and/or direct antiviral activity that restricts viral replication and antigen load/antigenic stimulation. Although some of the genes such as *DDX58*, *ADAR* and *OAS1* (and even genetic variants: *OAS1* rs10774671) were found to be associated with variations in both rubella vaccine [14–16] and measles vaccine-induced immune responses in our studies, most of the identified genetic variants and association patterns appear to be virus-specific, possibly due to differences between viruses with respect to their ability to induce and/or counter some of the involved proteins and antiviral pathways [11–13,58].

A major strength of our study is the use of a relatively large racially diverse cohort of healthy vaccines after two doses of MMR vaccine, with extensive demographic and clinical information available, and no known contact with wild type measles virus infection. This allowed us to probe for race-specific associations (in the two most prevalent racial groups, Caucasians and African-Americans) with vaccine-induced immunity outcomes in our study cohort. A novel aspect of the present study is the comprehensive list of genetic variants, genotyped across 12 key antiviral genes (never before studied in the context of measles vaccine response heterogeneity) and correlations with clinically relevant vaccine-induced immune outcomes, such as neutralizing antibody levels.

We recognize that there is a real possibility that some of the reported associations reflect false–positive results, particularly in the smaller sample sizes (and reduced power) available in our race-specific analysis. In our work we choose to report the findings for the two most prevalent racial groups (Caucasians and

African-Americans), separately, as the allele frequencies can vary considerably between racial and ethnic groups and thus can confound the results. Importantly, most of the genes/SNPs and genetic association patterns identified in our study were different between the Caucasian and African-American racial groups we studied, which is not surprising given the differences in allele frequencies for immune response genes reported for these two races [59], as well as the differences in allele/haplotype frequencies observed for the two groups in our study. It is plausible that different genetic polymorphisms would be important in other populations, such as African children, and our findings from a cohort of older children in Minnesota (an area with high vaccination coverage) may not apply to other areas of the world (with low vaccination coverage) where mainly younger children of different race/ethnicity contract the disease.

We also acknowledge the relatively large number of comparisons in our analyses and addressed the multiple testing issues by adjusting for FDR and supplementing the *p*-values with corresponding *q*-values. After correction for FDR, 15 SNPs (out of 32 reported single-SNP associations at the level $p \leq 0.01$; 11 SNPs in Caucasians and 4 SNPs in African-Americans) still remained significant at the *q*-value < 0.20 [38]. As this is the first study assessing polymorphisms in antiviral genes and their possible effect on measles vaccine-induced immunity in a race-specific manner, we chose to also use the less stringent *q*-value cutoff value of 0.5 (roughly corresponding to *p*-values ≤ 0.01 for single-SNP associations) believing that the risk of false negatives outweighs that of false positives.

Nevertheless, the high biological plausibility for some of the genetic variants, the haplotype-level associations, the observed allele dose-related relationships and cross-regulation of different immune outcomes by the same genetic variants, and considerable effect on immune outcome measures, increase our confidence in the reported findings. Follow-up replication and function-based studies are planned to validate and extend our results, and increase our understanding of the mechanisms, involved in immune response variations after measles vaccination. This will potentially inform the development of new and/or improved vaccines and support a novel paradigm of vaccine development, a concept we have called vaccinomics [9].

Taken together, the increasing evidence for the role of novel innate immunity genes (such as *DDX58*/*RIG-I*, *OAS* and *ADAR*), in the regulation of measles vaccine adaptive immunity will likely open new avenues for designing better vaccines and improved vaccination strategies.

Acknowledgements

We thank the parents and children who participated in our studies and the Mayo Vaccine Research Group nurses for subject recruitment. We thank the Mayo Vaccine Research Group laboratory personnel for technical help with the assays and Matthew J. Phan and Caroline Vitse for assistance in preparing the manuscript. We thank also Megan O'Byrne for her contribution to statistical analyses and Rick Kennedy for the useful discussions. We thank David Rider and Ying Li for developing the SNP selection algorithm, and Julie M. Cunningham and the Mayo Advanced Genomic Technology Center for assistance with genotyping. The project described was supported by Award Numbers AI33144 and AI48793 (which recently received a MERIT Award) from the National Institute Of Allergy And Infectious Diseases, and 5UL1RR024150-03 from the National Center for Research Resources (NCRR), a component of the National Institutes of Health, and the NIH Roadmap for Medical Research. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute Of Allergy And Infectious Diseases or the National Institutes of Health.

Disclosures: Dr. Poland is the chair of a DMSB for novel non-measles vaccines undergoing clinical study by Merck Research Laboratories. Dr. Jacobson recently served on a Safety Review Committee for a post-licensure study conducted by Kaiser-Permanente concerning Gardasil HPV vaccine funded by Merk & Co. Other authors do not have any conflicts of interest.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.vaccine.2011.09.043.

References

[1] Global measles mortality, 2000–2008. MMWR Morb Mortal Wkly Rep 2009;58(December (47)):1321–6.

[2] Plemper RK, Snyder JP. Measles control – can measles virus inhibitors make a difference? Curr Opin Investig Drugs 2009;10(8):811–20.

[3] Sugerman DE, Barskey AE, Delea MG, Ortega-Sanchez IR, Bi D, Ralston KJ, et al. Measles outbreak in a highly vaccinated population, San Diego, 2008: role of the intentionally undervaccinated. Pediatrics 2010;125(4):747–55.

[4] Jick H, Hagberg KW. Measles in the United Kingdom 1990–2008 and the effectiveness of measles vaccines. Vaccine 2010;28(29):4588–92.

[5] Kennedy AM, Gust DA. Measles outbreak associated with a church congregation: a study of immunization attitudes of congregation members. Public Health Rep 2008;123(2):126–34.

[6] Measles – United States, January 1–April 25, 2008. MMWR Morb Mortal Wkly Rep 2008;57(18):494–8.

[7] Orenstein WA, Strebel PM, Hinman AR. Building an immunity fence against measles. J Infect Dis 2007;196(10):1433–5.

[8] Poland GA, Ovsyannikova IG, Jacobson RM, Smith DI. Heterogeneity in vaccine immune response: the role of immunogenetics and the emerging field of vaccinomics. Clin Pharmacol Ther 2007;82(6):653–64.

[9] Poland GA, Ovsyannikova IG, Kennedy RB, Haralambieva IH, Jacobson RM. Vaccinomics and a new paradigm for the development of preventive vaccines against viral infections. Omics 2011;15(9):625–36.

[10] Samuel CE. Antiviral actions of interferons. Clin Microbiol Rev 2001;14(4):778–809.

[11] Garcia-Sastre A, Biron CA. Type 1 interferons and the virus-host relationship: a lesson in detente. Science 2006;312(5775):879–82.

[12] Randall RE, Goodbourn S. Interferons and viruses: an interplay between induction, signalling, antiviral responses and virus countermeasures. J Gen Virol 2008;89(Pt 1):1–47.

[13] Sadler AJ, Williams BR. Interferon-inducible antiviral effectors. Nat Rev Immunol 2008;8(7):559–68.

[14] Haralambieva IH, Dhiman N, Ovsyannikova IG, Vierkant RA, Pankratz VS, Jacobson RM, et al. 2′–5′-Oligoadenylate synthetase single-nucleotide polymorphisms and haplotypes are associated with variations in immune responses to rubella vaccine. Hum Immunol 2010;71(4):383–91.

[15] Ovsyannikova IG, Haralambieva IH, Dhiman N, O'Byrne MM, Pankratz VS, Jacobson RM, et al. Polymorphisms in the vitamin A receptor and innate immunity genes influence the antibody response to rubella vaccination. J Infect Dis 2010;201(2):207–13.

[16] Ovsyannikova IG, Dhiman N, Haralambieva IH, Vierkant RA, O'Byrne MM, Jacobson RM, et al. Rubella vaccine-induced cellular immunity: evidence of associations with polymorphisms in the Toll-like, vitamin A and D receptors, and innate immune response genes. Hum Genet 2010;127:207–21.

[17] Hijikata M, Ohta Y, Mishiro S. Identification of a single nucleotide polymorphism in the M × A gene promoter (G/T at nt -88) correlated with the response of hepatitis C patients to interferon. Intervirology 2000;43(2):124–7.

[18] Suzuki F, Arase Y, Suzuki Y, Tsubota A, Akuta N, Hosaka T, et al. Single nucleotide polymorphism of the M × A gene promoter influences the response to interferon monotherapy in patients with hepatitis C viral infection. J Viral Hepat 2004;11(3):271–6.

[19] Torisu H, Kusuhara K, Kira R, Bassuny WM, Sakai Y, Sanefuji M, et al. Functional M × A promoter polymorphism associated with subacute sclerosing panencephalitis. Neurology 2004;62(3):457–60.

[20] King JK, Yeh SH, Lin MW, Liu CJ, Lai MY, Kao JH, et al. Genetic polymorphisms in interferon pathway and response to interferon treatment in hepatitis B patients: a pilot study. Hepatology 2002;36(6):1416–24.

[21] Yakub I, Lillibridge KM, Moran A, Gonzalez OY, Belmont J, Gibbs RA, et al. Single nucleotide polymorphisms in genes for 2′–5′-oligoadenylate synthetase and RNase L inpatients hospitalized with West Nile virus infection. J Infect Dis 2005;192(10):1741–8.

[22] Knapp S, Yee LJ, Frodsham AJ, Hennig BJ, Hellier S, Zhang L, et al. Polymorphisms in interferon-induced genes and the outcome of hepatitis C virus infection: roles of M × A, OAS-1 and PKR. Genes Immun 2003;4(6):411–9.

[23] Hamano E, Hijikata M, Itoyama S, Quy T, Phi NC, Long HT, et al. Polymorphisms of interferon-inducible genes OAS-1 and M × A associated with SARS in the Vietnamese population. Biochem Biophys Res Commun 2005;329(4):1234–9.
[24] He J, Feng D, de Vlas SJ, Wang H, Fontanet A, Zhang P, et al. Association of SARS susceptibility with single nucleic acid polymorphisms of OAS1 and M × A genes: a case–control study. BMC Infect Dis 2006;6:106–12.
[25] Haralambieva IH, Ovsyannikova IG, O'Byrne M, Pankratz VS, Jacobson RM, Poland GA. A large observational study to concurrently assess persistence of measles specific B-cell and T-cell immunity in individuals following two doses of MMR vaccine. Vaccine 2011;29(27):4485–91.
[26] Ovsyannikova IG, Haralambieva IH, Vierkant RA, Pankratz VS, Poland GA. The role of polymorphisms in Toll-like receptors and their associated intracellular signaling genes in measles vaccine immunity. Hum Genet 2011;(Mar), in press, doi:10.1007/s00439-011-0977-x.
[27] Haralambieva IH, Ovsyannikova IG, Vierkant RA, Poland GA. Development of a novel efficient fluorescence-based plaque reduction microneutralization assay for measles immunity. Clin Vaccine Immunol 2008;15(7):1054–9.
[28] Ryan JE, Ovsyannikova IG, Dhiman N, Pinsky NA, Vierkant RA, Jacobson RM, et al. Inter-operator variation in ELISPOT analysis of measles virus-specific IFN-gamma secreting T cells. Scand J Clin Lab Invest 2005;65(8):681–90.
[29] Ovsyannikova IG, Dhiman N, Jacobson RM, Vierkant RA, Pankratz VS, Poland GA. HLA homozygosity does not adversely effect measles vaccine-induced cytokine responses. Virology 2007;364(1):87–94.
[30] Ryan JE, Dhiman N, Ovsyannikova IG, Vierkant RA, Pankratz VS, Poland GA. Response surface methodology to determine optimal cytokine responses in human peripheral blood mononuclear cells after smallpox vaccination. J Immunol Methods 2009;341(1–2):97–105.
[31] Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, Nickerson DA. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. Am J Hum Genet 2004;74(1):106–20.
[32] den Dunnen JT, Antonarakis SE. Nomenclature for the description of human sequence variations. Hum Genet 2001;109(1):121–4.
[33] Weir BS. Genetic data Analysis II: methods for discrete population genetic data. Sinauer Associates, Inc; 1996. p. 98–9.
[34] Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 2005;21(2):263–5.
[35] Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA. Score tests for association between traits and haplotypes when linkage phase is ambiguous. Am J Hum Genet 2002;70:425–34.
[36] Storey JD. A direct approach to false discovery rates. J R Statist Soc B 2002;64:479–98.
[37] Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci USA 2003;100(16):9440–5.
[38] Dhiman N, Haralambieva IH, Kennedy RB, Vierkant RA, O'Byrne MM, Ovsyannikova IG, et al. SNP/haplotype associations in cytokine and cytokine receptor genes and immunity to rubella vaccine. Immunogenetics 2010;62(4):197–210.
[39] Lim JK, Lisco A, McDermott DH, Huynh L, Ward JM, Johnson B, et al. Genetic variation in OAS1 is a risk factor for initial infection with West Nile virus in man. PLoS Pathogens 2009;5(2):e1000321.
[40] Dhiman N, Poland GA, Cunningham JM, Jacobson RM, Ovsyannikova IG, Vierkant RA, et al. Variations in measles vaccine-specific humoral immunity by polymorphisms in SLAM and CD46 measles virus receptors. J Allergy Clin Immunol 2007;120(3):666–72.
[41] Clifford HD, Richmond P, Khoo SK, Zhang G, Yerkovich ST, Le Souëf PN, et al. SLAM and DC-SIGN measles receptor polymorphisms and their impact on antibody and cytokine responses to measles vaccine. Vaccine 2011;29(33):5407–13.
[42] Ovsyannikova IG, Haralambieva IH, Vierkant RA, O'Byrne MM, Jacobson RM, Poland GA. The association of CD46, SLAM, and CD209 cellular receptor gene SNPs with variations in measles vaccine-induced immune responses – a replication study and examination of novel polymorphisms. Hum Hered 2011;(Aug), in press.
[43] Dhiman N, Ovsyannikova IG, Vierkant RA, Ryan JE, Pankratz VS, Jacobson RM, et al. Associations between SNPs in toll-like receptors and related intracellular signaling molecules and immune responses to measles vaccine: preliminary results. Vaccine 2008;26(14):1731–6.
[44] Dhiman N, Ovsyannikova IG, Cunningham JM, Vierkant RA, Kennedy RB, Pankratz VS, et al. Associations between measles vaccine immunity and single nucleotide polymorphisms in cytokine and cytokine receptor genes. J Infect Dis 2007;195(1):21–9.
[45] Haralambieva IH, Ovsyannikova IG, Kennedy RB, Vierkant RA, Pankratz VS, Jacobson RM, et al. Associations between single nucleotide polymorphisms and haplotypes in cytokine and cytokine receptor genes and immunity to measles vaccination. Vaccine, 2011;(Aug), in press, doi:10.1016/j.vaccine.2011.08.083.
[46] Yucesoy B, Johnson VJ, Fluharty K, Kashon ML, Slaven JE, Wilson NW, et al. Influence of cytokine gene variations on immunization to childhood vaccines. Vaccine 2009;27(50):6991–7.
[47] Haralambieva IH, Ovsyannikova IG, Dhiman N, Kennedy RB, O'Byrne M, Pankratz VS, et al. Common SNPs/Haplotypes in IL18R1 and IL18 genes are associated with variations in humoral immunity to smallpox vaccination in caucasians and African-Americans. J Infect Dis 2011;204(3):433–41.
[48] Dhiman N, Ovsyannikova IG, Vierkant RA, Pankratz VS, Jacobson RM, Poland GA. Associations between cytokine/cytokine receptor SNPs and humoral immunity to measles, mumps and rubella in a Somali population. Tissue Antigens 2008;72(3):211–20.
[49] Moore CB, Ting JP. Regulation of mitochondrial antiviral signaling pathways. Immunity 2008;28(6):735–9.
[50] Yoneyama M, Fujita T. Function of RIG-I-like receptors in antiviral innate immunity. J Biol Chem 2007;282(21):15315–8.
[51] Benn CS, Aaby P, Bale C, Olsen J, Michaelsen KF, George E, et al. Randomised trial of effect of vitamin A supplementation on antibody response to measles vaccine in Guinea-Bissau, west Africa. Lancet 1997;350(9071):101–5.
[52] Soye KJ, Trottier C, Richardson CD, Ward BJ, Miller Jr WH. RIG-I is required for the inhibition of measles virus by retinoids. PLoS ONE 2011;6(7):e22323.
[53] Ikegame S, Takeda M, Ohno S, Nakatsu Y, Nakanishi Y, Yanagi Y. Both RIG-I and MDA5 RNA helicases contribute to the induction of alpha/beta interferon in measles virus-infected human cells. J Virol 2010;84(1):372–9.
[54] Hu J, Nistal-Villan E, Voho A, Ganee A, Kumar M, Ding Y, et al. A common polymorphism in the caspase recruitment domain of RIG-I modifies the innate immune response of human dendritic cells. J Immunol 2010;185(1):424–32.
[55] Bonnevie-Nielsen V, Field LL, Lu S, Zheng DJ, Li M, Martensen PM, et al. Variation in antiviral 2′,5′-oligoadenylate synthetase (2′5′AS) enzyme activity is controlled by a single-nucleotide polymorphism at a splice-acceptor site in the OAS1 gene. Am J Hum Genet 2005;76(4):623–33.
[56] Yang JH, Nie Y, Zhao Q, Su Y, Pypaert M, Su H, et al. Intracellular localization of differentially regulated RNA-specific adenosine deaminase isoforms in inflammation. J Biol Chem 2003;278(46):45833–42.
[57] Umlauf BJ, Ovsyannikova IG, Haralambieva IH, Kennedy RB, Vierkant RA, Pankratz VS, et al. Correlations between vaccinia-specific immune responses within a cohort of armed forces members. Viral Immunol 2011;24(5), in press, doi:10.1089/vim.2011.0029.
[58] Pichlmair A, Reis e Sousa C. Innate recognition of viruses. Immunity 2007;27(3):370–83.
[59] Martin AM, Athanasiadis G, Greshock JD, Fisher J, Lux MP, Calzone K, et al. Population frequencies of single nucleotide polymorphisms (SNPs) in immunomodulatory genes. Hum Hered 2003;55(4):171–8.

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/51785748

Genetics of migraine in the age of genome-wide association studies

Article *in* The Journal of Headache and Pain · November 2011

DOI: 10.1007/s10194-011-0399-0 · Source: PubMed

CITATIONS
96

READS
84

1 author:

Markus Schürks
Bayer
137 PUBLICATIONS **9,683** CITATIONS

SEE PROFILE

All content following this page was uploaded by Markus Schürks on 16 May 2014.

The user has requested enhancement of the downloaded file.

J Headache Pain (2012) 13:1–9
DOI 10.1007/s10194-011-0399-0

REVIEW ARTICLE

Genetics of migraine in the age of genome-wide association studies

Markus Schürks

Received: 21 September 2011 / Accepted: 24 October 2011 / Published online: 11 November 2011
© The Author(s) 2011. This article is published with open access at Springerlink.com

Abstract Genetic factors importantly contribute to migraine. However, unlike for rare monogenic forms of migraine, approaches to identify genes for common forms of migraine have been of limited success. Candidate gene association studies were often negative and positive results were often not replicated or replication failed. Further, the significance of positive results from linkage studies remains unclear owing to the inability to pinpoint the genes under the peaks that may be involved in migraine. Problems hampering these studies include limited sample sizes, methods of migraine ascertainment, and the heterogeneous clinical phenotype. Three genome-wide association studies are available now and have successfully identified four new genetic variants associated with migraine. One new variant (rs1835740) modulates glutamate homeostasis, thus integrates well with current concepts of neurotransmitter disturbances. This variant may be more specific for severe forms of migraine such as migraine with aura than migraine without aura. Another variant (rs11172113) implicates the lipoprotein receptor LRP1, which may interact with neuronal glutamate receptors, thus also providing a link to the glutamate pathway. In contrast, rs10166942 is in close proximity to *TRPM8*, which codes for a cold and pain sensor. For the first time this links a gene explicitly implicated in pain related pathways to migraine. The potential function of the fourth variant rs2651899 (*PRDM16*) in migraine is unclear. All these variants only confer a small to moderate change in risk for migraine, which concurs with migraine being a heterogeneous disorder. Ongoing large international collaborations will likely identify additional gene variants for migraine.

Keywords Migraine · Migraine with aura · Genetics · Genome-wide association studies · Glutamate · *TRPM8* · *PRDM16* · *LRP1*

Introduction

Migraine is a common neurological disorder affecting between 10 and 20% of the population [1]. The clinical presentation is heterogeneous and includes recurrent headache attacks, associated symptoms of vegetative disturbance, and hypersensitivity of various functional systems of the nervous system [1, 2]. About one-third of migraineurs further experience transient neurological symptoms mostly involving the visual system prior to or during a migraine attack, which are known as migraine aura [2]. The International Headache Society (IHS) has established gold-standard criteria for the diagnosis of migraine [3]. While migraines with aura (MA) and without aura (MO) are the most common forms, the IHS also acknowledges, for example, certain forms lacking one diagnostic criterion as probable migraine as well as rare familial monogenetic forms.

Heredity has been shown to play an important role in migraine pathogenesis. About 50% of affected individuals have a first-degree relative also suffering from migraine [4–6]. In addition, family and twin studies support the idea of MO and MA being different phenotypes of the same entity, with a heritability ranging from 33 to 57% [5, 7, 8].

The broad clinical spectrum of migraine and the results from heritability studies have posted a challenge to the

M. Schürks (✉)
Department of Neurology, University Hospital Essen,
Hufelandstrasse 55, 45122 Essen, Germany
e-mail: markus.schuerks@uni-due.de

understanding of the underlying pathophysiology and molecular genetics. For example, it is difficult to find a pathophysiological link between a headache, which is not symptomatic, and symptoms like nausea, photophobia, or hypotension. In addition, it is unclear if some of the clinical symptoms are more genetically determined and some are more environmentally determined. On the contrary, dissecting the migraine phenotype into endophenotypes has offered opportunities to better understand the pathophysiological processes involved, facilitating the formulation of hypotheses for genetic studies. For example, symptoms like yawning (premonitory phase) and nausea (premonitory and headache phase) point toward an involvement of dopamine and have entailed genetic association studies of variants in genes involved in dopamine metabolism [9]. Another approach was to utilize epidemiological data for hypothesis generation. The three- to fourfold higher prevalence of migraine among women than men [10] as well as attacks following a menstrual cycle pattern [11] suggested an involvement of sex hormones and led to the investigation of genetic variants in sex hormone receptor genes [12]. Aware of the limitations of these approaches, methods free of assumptions have been employed mining the genome for associations with the migraine phenotype or certain migraine symptoms. These include linkage studies and genome-wide association studies (GWAS). In contrast to all approaches focusing on common forms of migraine, investigations in rare familial forms of migraine, presenting with a well circumscribed aura in the form of a hemiparesis/hemiplegia (familial hemiplegic migraine, FHM), have been more successful in identifying underlying genetic causes [13–15].

This manuscript does not aim at providing a complete overview of genetic studies in migraine; this has comprehensively been done recently [16, 17]. This manuscript will provide examples for different approaches used in migraine genetics and then focus on results from the recently published GWASs.

Familial hemiplegic migraine (FHM)

FHM is a rare, but severe form of migraine displaying an autosomal dominant mode of transmission. The main characteristic of FHM is a prolonged aura in the form of a hemiparesis/hemiplegia presenting with the migraine headache. FHM belongs to the group of channelopathies and multiple genetic variants in three different genes coding for neuronal ion channels have been identified as causative. These genetic variants perturb the activity of the channels, thus increasing neuronal excitability. In FHM type 1 the affected gene is *CACNA1A* (chromosome 19p13), coding for a P/Q calcium channel [15], in FHM type 2 the *ATP1A2* gene (chromosome 1q23) is affected, encoding a subunit of the Na^+/K^+-ATPase [13], and FHM type 3 is caused by variants in the *SCN1A* gene (chromosome 2q24), coding for a sodium channel [14]. However, patients with FHM only constitute a very small proportion of all migraineurs and there is little evidence, that the identified genes play a major role in common forms of migraine [18]. This may also suggest that the ion channel genes are only important for the aura and not the migraine per se, since, in particular, the *CACNA1A* gene is also implicated in other neurological disorders not associated with migraine including episodic ataxia type 2 and spinocerebellar ataxia type 6.

Candidate gene approaches

Candidate gene approaches have been a popular tool in studying migraine genetics. Genes investigated are, for example, implicated in serotonin and dopamine metabolism, neurogenic inflammation, vascular function, and hormone regulation.

Migraine is viewed as an inherited brain disorder, characterized by neurotransmitter imbalances that lead to neuronal dysfunctions [19]. The serotonergic system plays an important role and alterations in serotonin metabolism and in the processing of central serotonin-mediated responses are typical features among migraineurs [20]. Following axonal release serotonin (5-HT) is rapidly taken up again by the presynaptic serotonin transporter SLC6A4 (5-HTT) [21]. An insertion/deletion polymorphism in the promoter region of the SLC6A4 gene, termed 5-HTTLPR [22, 23] has functional impact on the synaptic clearing of serotonin [24]. Ten studies in different populations have investigated the association between the *5-HTTLPR* gene and migraine among Europeans and Asians [25]. Pooled analyses did not find an overall association; however, there was some suggestion that gender and migraine aura status may have modifying roles among Europeans.

Another neurotransmitter implicated in migraine pathogenesis is dopamine. Symptoms like yawning in the premonitory phase and nausea in both the premonitory and headache phase are ascribed to dopamine [26]. There is also evidence from other pain disorders that dopamine plays a role in central pain [27]. Further support for a role of dopamine in migraine comes from studies showing that anti-dopaminergic drugs exert beneficial effects, when administered during migraine attacks and from experimental work showing that dopamine modulates the neuronal firing rate in the trigeminocervical complex [9]. Genetic association studies have mostly investigated variants in dopamine receptor genes and the dopamine beta-hydroxylase gene; however, study results are conflicting and the significance of these variants in migraine remains unclear.

A “neurogenic inflammation” appears to be a key mechanism in causing the migraine headache by activating the trigeminal system [28]. Tumor necrosis factor-alpha (TNF-alpha) is an important inflammatory cytokine and modulator of immune responses. Its role in migraine has been postulated based on reports of TNF-alpha concentration changes in serum [29–31] and urine [32] among migraineurs, as well as findings that TNF-alpha can stimulate transcription of calcitonin gene-related peptide (CGRP), which is pivotal in migraine pathophysiology [33]. Variants in the genes coding for TNF-alpha and the closely related TNF-beta have been shown to modulate cytokine levels of TNF-alpha and TNF-beta [34, 35] prompting investigations of various polymorphisms in these genes with respect to their role in migraine. Among the ten studies investigating variants in these genes, the best evidence is available for the *TNF-alpha* -308G>A and *TNF-beta* 252A>G polymorphisms. Pooled analyses did not indicate an overall association with migraine [36]. While subgroup analyses suggested that associations may differ among populations of non-Caucasian origin, these findings only stem from few studies.

Migraine prevalence is three- to fourfold higher in women than men [10]. Further, some women suffer from migraine attacks presenting at specific times during the menstrual cycle or experience changing patterns of attacks during pregnancy or after menopause [11, 37]. Finally, changes in estrogen levels can trigger migraine attacks [11, 37]. These findings have prompted studies investigating variants in genes coding for proteins in sex hormone receptor pathways and metabolism among migraineurs. Pooled analyses of the available studies suggest that the estrogen receptor-1 gene (*ESR-1*) 594 G>A and 325 C>G polymorphisms can modify the risk for migraine, a finding that does not differ between MA and MO [12]. In contrast, the *ESR-1* Pvu II C>T polymorphism and the progesterone receptor (*PGR*) PROGINS insert polymorphism do not appear to be associated with migraine.

The vasculature appears to play a special role in migraine. For example, migraineurs carry an approximately twofold increased risk for ischemic stroke compared to non-migraineurs [38], some vasculopathies are frequently observed among migraineurs such as livedo reticularis [39] and Sneddon’s syndrome [40], and altered vascular reactivity can be found among young migraineurs even in the absence of other disorders [41]. These findings may be linked to oxidative stress causing endothelial dysfunction [42]. Specifically, certain gene variants in the methylenetetrahydrofolate reductase gene (*MTHFR* 677C>T polymorphism) and angiotensin-converting enzyme gene (*ACE* D/I polymorphism) appear to importantly contribute to the vascular oxidative stress response [42]. Consequently, an association between the *MTHFR* 677C>T and *ACE* D/I polymorphisms with migraine has been suggested [42]. Both are functional variants. The *MTHFR* 677TT genotype impairs enzyme activity and carriers have increased homocysteine levels. Carriers of the *ACE* II genotype have ACE plasma levels half that of DD subjects, with ID subjects having intermediate levels. Association studies investigating the role of these variants in migraine were controversial. Pooled analyses of 13 studies indicated that the *MTHFR* 677TT genotype was associated with a 48% increased risk for MA (pooled odds ratio [OR] = 1.48, 95% confidence interval [CI] 1.02–2.13), but not MO [43]. In contrast, pooled results from nine studies indicated that the *ACE* II genotype was associated with a reduced risk for MA (pooled OR = 0.71, 95% CI 0.55–0.93) and MO (pooled OR = 0.84, 95% CI 0.70–0.99) [43].

Genome-wide approaches

In contrast to candidate gene association studies, genome-wide approaches have been employed to identify potential new chromosomal regions or genes that may associate with migraine, migraine subtypes, or migraine features. These are basically data mining techniques screening a part or the whole genome. The two main methods employed are linkage analyses that study pedigrees of migraineurs and GWAS that investigate large populations of migraineurs and non-migraineurs.

Linkage studies

Linkage analyses have identified several genetic loci for overall migraine, MA, and MO. However, the underlying genes from these chromosomal regions that may plausibly alter risk of migraine have not yet been identified. Overall migraine has been linked to chromosomes Xq24-28 [44], 6p12.2-21.1 [45], 19p13 [46], and 1q31 [47]; MA to 4q24 [48], 9q21-q22 [49], 11q24 [50], and 15q11-13 [51]; and MO to 14q21.1-22.3 [52], and 4q21 [53].

Alternative phenotyping methods

Only two independent linkage studies from Iceland and Finland describe an overlapping gene locus on chromosome 4 [48, 53]. However, the other studies have reported many different loci, which is likely a reflection of the genetic heterogeneity of migraine. The heterogeneous migraine phenotype with a broad spectrum poses a particular challenge to genetic studies, which may in part explain the genetic heterogeneity observed. Hence, more clearly defined individual features of migraine (“endophenotypes”) may help to more consistently identify genetic loci or genes involved in migraine [54]. For example, employing trait component analysis and latent class analysis has successfully linked certain chromosomal regions to pulsating pain, phono-/photophobia, nausea, and age at onset of migraine [44, 55, 56].

Genome-wide association studies

The first GWAS in migraine was published by the International Headache Genetics Consortium (IHGC) in 2010 [57] (Table 1). This was a two-stage association study involving seven clinic-based case collections. In the discovery stage 2,731 migraineurs all suffering from MA, recruited from clinics in Finland, Germany, and the Netherlands, were compared with 10,747 population-based controls from the same countries. In the replication stage an additional 3,202 cases from clinics in Denmark, the Netherlands, and Germany as well as from the general population in Iceland were available and compared to 40,062 population-based controls. Patients used for replication suffered either from MA or MO. Diagnoses were made by headache experts using a combination of questionnaires and direct interviews based on the 2004 IHS diagnostic criteria [3]. The following diagnostic subgroups were analyzed: overall migraine, MA only, MA and MO, and MO only.

In the discovery stage, 429,912 genetic variants were genotyped in cases and controls utilizing an Illumina array. Considering a genome-wide significance threshold of $p \leq 5 \times 10^{-8}$, only one single nucleotide polymorphism (SNP; rs1835740) was significantly associated with overall migraine. When using a more liberal threshold of $p \leq 5 \times 10^{-5}$ eleven SNPs were associated with migraine. The minor allele of rs1835740 was associated with an increased risk for migraine (OR = 1.23, 95% CI 1.15–1.32, $p = 5.38 \times 10^{-9}$). The population attributable risk for this marker was estimated to be 10.7%. Replication of the association at rs1835740 was successful and the meta-analysis of results from all seven case–control samples indicated an 18% increased risk for overall migraine (OR = 1.18, 95% CI 1.13–1.24, $p = 1.69 \times 10^{-11}$). In subgroup analyses, the effects were stronger for patients with MA than in those suffering from MO.

Findings from this GWAS lend support for a role of glutamate in the pathophysiology of migraine, in particular MA. rs1835740 is located between the genes *MTDH* and *PGCP*, which are involved in glutamate homeostasis. The study authors have also shown that the minor allele of rs1835740 increases the expression levels of *MTDH* [57]. Further, previous studies in astrocytes have indicated that *MTDH* downregulates *SLC1A2*, which codes for the major glutamate transporter in the brain. This may lead to an increase in the extracellular concentration of glutamate, the major excitatory transmitter in the central nervous system. Glutamate plays a role in trigeminovascular pain processing [58] and may modulate the threshold for cortical spreading depression [59], the putative biological correlate of the aura [60].

In addition to rs1835740, eleven SNPs were associated with migraine with $p \leq 5 \times 10^{-5}$ (Table 1). Although missing genome-wide significance, these SNPs may have potential implications for migraine. Nine of these are located within or close to known genes including *SMYD3*, *INSIG2*, *TRPM8*, *MYLK4*, *ZNF311*, *SGCZ*, *NAV2*, *COG3*, and *AQP9*.

The second GWAS was performed by the Dutch-Icelandic (DICE) migraine genetics consortium [61] (Table 1). In the discovery stage, five Dutch and one Icelandic population-based cohorts with 2,446 migraineurs and 8,534 non-migraineurs were investigated. The replication samples derived from two population-based Dutch cohorts and one Australian Twin study totaling 2,957 migraineurs and 5,774 non-migraineurs. Migraine ascertainment in the discovery cohorts was primarily by means of questionnaires, and genotyping was performed utilizing a variety of genotyping platforms. No SNP was associated with migraine at the genome-wide level, 32 SNPs were associated with migraine at $p < 10^{-5}$. The best association was seen for rs9908234 ($p = 8.00 \times 10^{-8}$), which is located in the nerve growth factor (NGF) gene. However, neither this variant nor 18 other chosen variants out of the 32 variants showed associations in the replication cohorts. In a further step, the authors explored associations of previously identified candidate gene variants for migraine with migraine in the pooled discovery dataset. None of the candidates was associated with migraine; however, in gene-based analyses there was modest support for an association of *MTDH*, implicated in the previous GWAS [57], with migraine. Among the reasons potentially explaining the inability to replicate findings from the discovery cohorts are usage of different genotyping platforms, methods of migraine ascertainment, and a lack of power, despite the efforts of bringing together many cohorts.

The third GWAS was also performed to identify common genetic variants for migraine at the population level [62] (Table 1). The study included three population-based cohorts and information from the IHGC. The population-based cohorts were the Women’s Genome Health Study (WGHS), the Genetic Epidemiology of Migraine (GEM) study, and the Study of Health in Pomerania (SHIP). Migraine was either self-reported on questionnaires, diagnosed by headache experts or by a combination of both.

Table 1 Overview of published GWAS in migraine

Study	Anttila et al. [57]	Ligthart et al. [61]	Chasman et al. [62]
Consortium name	IHGC	DICE	(–)
Discovery samples			
Names	?,?, LUMINA	ERF, NESDA, NTR1, NTR2, Rotterdam study, AGES study	WGHS
Origin	Clinic-based; Finland, Germany, The Netherlands	Population-based; The Netherlands, Iceland	Population-based; US, European descent
# of migraineurs/ non-migraineurs	2,731/10,747	2,446/8,534	5,122/18,108
Implicated gene variants (*gene*)[a]	rs1835740 (between *MTDH* and *PGCP*)	rs9908234 (*NGFR*)	rs2651899 (*PRDM16*) rs2078371 (–) rs10166942 (*TRPM8*) rs17172526 (*SEPT7*) rs2203834 (*C8orf79*) rs13290757 (–) rs11172113 (*LRP1*)
Replication samples			
Names	?, ?, ?, LUMINA	GEM, NTR3, ATM	GEM, SHIP, IHGC
Origin	Clinic- and population-based/Denmark, Iceland, Germany, The Netherlands	Population-based/The Netherlands, Australia	Population-based/The Netherlands, Germany; clinic-based/Europe
# of migraineurs/ non-migraineurs	3,202/40,062	2,957/5,774	3,828/13,949
Replicated gene variants	rs1835740 (between *MTDH* and *PGCP*)	None	rs2651899 (*PRDM16*) rs10166942 (*TRPM8*) rs11172113 (*LRP1*)
Meta-analysis among	Discovery and replication samples	Discovery samples	Discovery and replication samples
Pooled results[a]	rs1835740: OR = 1.18, 95% CI 1.13–1.24, $p = 1.69 \times 10^{-11}$	rs9908234: $p = 8.00 \times 10^{-8}$	rs2651899: OR = 1.11, 95% CI 1.07–1.15, $p = 3.8 \times 10^{-9}$ rs10166942: OR = 0.85, 95% CI 0.82–0.89, $p = 5.5 \times 10^{-12}$ rs11172113: OR = 0.90, 95% CI 0.87–0.93, $p = 4.3 \times 10^{-9}$
Other implicated gene variants in discovery samples			
Significance threshold	$p \leq 5 \times 10^{-5}$	$p < 10^{-5}$	$p < 5 \times 10^{-6}$
Gene variants (*genes*)[a]	rs12084862 (*SMYD3*) rs17528324 (*INSIG2*) rs17862920 (*TRPM8*) rs2038761 (*MYLK4*) rs6456880 (*ZNF311*) rs7753655 (–) rs10888075 (near *SGCZ*) rs10111769 (–) rs2042600 (*NAV2*) rs3794331 (*COG3*) rs473422 (near *AQP9*)	31 other SNPs including: rs11636768 (*AGBL1*) rs10275320 (*MACC1*) rs4939879 (*LIPG*) rs4861775 (*AGA*) rs986222 (*KIF2OB*) rs6107848 (*BMP2*) rs140174 (*IGLL1*) rs1146161 (*TSPAN2*) rs4742323 (*KDM4C*)	rs2078371 (–) rs17172526 (*SEPT7*) rs2203834 (*C8orf79*) rs13290757 (–)

IHGC International Headache Genetics consortium, *LUMINA* Leiden University Migraine Neuro Analysis, *DICE* Dutch-Icelandic migraine genetics consortium, *ERF* Erasmus Rucphen Family study, *NESDA* The Netherlands Study of Depression and Anxiety, *NTR* The Netherlands Twin Registry, *GEM* Genetic Epidemiology of Migraine, *ATM* Australian Twin Migraine GWA study, *WGHS* Women's Genome Health Study, *SHIP* Study of Health in Pomerania, *OR* odds ratio, *CI* confidence interval

[a] Results for overall migraine

Genotyping was performed by Illumina chips (WGHS, IHCG), Affymetrix arrays (SHIP), and iPLEX method from Sequenom (GEM).

The discovery cohort consisted of 23,230 women of European ancestry participating in the WGHS, which is a large population-based cohort in the US. Among those 5,122 women reported migraine, while 18,108 did not. In the genome-wide scan using age-adjusted logistic regression and assuming an additive genetic model, none of the investigated SNPs were associated with migraine at $p < 5 \times 10^{-8}$. However, seven independent genetic loci had at least one SNP with an association of $p < 5 \times 10^{-6}$. From each locus the authors selected the most significant SNP (Table 1). None of these seven SNPs was associated with non-migraine headache. In addition, none of the seven SNPs showed associations with migraine in non-additive models, and no other SNP in the genome-wide scan reached genome-wide significance in non-additive models. Finally, association analysis using imputed genotypes of about 2.6 million SNPs identified the same seven loci as found with the genotyped SNPs.

The seven chosen SNPs were evaluated in GEM (774 migraineurs, 942 non-migraineurs), SHIP (306 migraineurs, 2,260 non-migraineurs), and IHGC (2,748 migraineurs, 10,747 non-migraineurs). The effect estimates for three of the SNPs (rs2651899, rs10166942, rs11172113) pointed in the same direction and were comparable in magnitude in all three replication studies as in the WGHS. In the next step a meta-analysis of results for the seven primary SNPs from all four cohorts was performed using a fixed-effects model, since there was no strong evidence of heterogeneity. All three SNPs identified from the replication cohorts based on concordance in direction and magnitude of association were associated with migraine at the genome-wide level (rs2651899: OR = 1.11, 95% CI 1.07–1.15, $p = 3.8 \times 10^{-9}$; rs10166942: OR = 0.85, 95% CI 0.82–0.89, $p = 5.5 \times 10^{-12}$; rs11172113: OR = 0.90, 95% CI 0.87–0.93, $p = 4.3 \times 10^{-9}$). rs2651899 at 1p36.32 is located within the first intron of *PRDM16*, rs10166942 at 2q37 is very close to the transcription start site for *TRPM8*, and rs11172113 at 12q13.3 is in the first intron of *LRP1*. Interestingly *TRPM8* was associated with migraine in the first IHGC GWAS at the sub-genome-wide level [57].

Gender-stratified analyses suggested that the association seen at rs10166942 (*TRPM8*) may be specific for women (meta-regression on gender $p = 0.004$). However, this potential differential association did not appear to be related to genetic variants in the *ESR1* gene, as there was no interaction between rs10166942 (*TRPM8*) and *ESR1* variants. Further investigations looking at potential differences according to migraine aura status and features (for example, unilateral pain location, pulsating pain quality, photophobia, etc.) did not suggest that the associations seen for any of the three variants differed by the presence or the absence of aura or any of the features. Finally, the associations seen for rs2651899 (*PRDM16*) and rs10166942 (*TRPM8*), but not rs11172113 (*LRP1*), were significantly associated with migraine compared to non-migraine headache, underlining the specificity of these variants for migraine.

TRPM8 codes for a cold and cold-induced burning pain sensor [63] which is primarily expressed in sensory neurons and the dorsal root ganglion [64]. *TRPM8* is investigated as a target in animal models of neuropathic pain [65]. Since migraine and neuropathic pain share some characteristics [66], a role for *TRPM8* in migraine as well as a link between both pain syndromes is plausible.

LRP1 belongs to the lipoprotein receptor family and is ubiquitously expressed [67]. LRP1 and NMDA glutamate receptors co-localize on neurons and they interact. This is particularly interesting since the previous GWAS by the IHGC has identified a genetic variant involved in glutamate homeostasis [57]. Both findings underline the role of glutamate in migraine pathophysiology and integrate well with pharmacological approaches targeting glutamate receptors for migraine treatment [58].

A role for *PRDM16* in migraine is unknown. Originally identified at a chromosomal breakpoint associated with myelodysplastic syndrome and leukemia [68], more recent research has focused on its role in brown fat development [69].

Implications

Endeavors to decipher the genetics of migraine have been successful with respect to FHM; however, there is ongoing debate about whether the genes involved (all coding for ion channels) are specific for migraine headache, the accompanying hemiplegic aura, or possibly for both. In contrast, attempts for common forms of migraine, although multifold, were often ambiguous. Positive results from candidate gene association studies have mostly not been robust, with the exception, for example, for the *MTHFR* 677C<T polymorphism, which shows an association with migraine in pooled analysis of the available studies. Many other genetic variants were not associated with migraine or replication of positive results is either missing or failed. Linkage studies produced some encouraging results, but it remains unclear which genes under the linkage peaks may be important for migraine. Reasons hampering an unequivocal success of previous studies include limited sample size, heterogeneity of the migraine phenotype as well as study design and conduct.

Migraine GWASs have overcome some of these drawbacks and advanced our understanding of migraine genetics with the following implications. First, available GWASs have identified four new genetic variants that are associated with migraine with genome-wide significance and none of the implicated genes has been a candidate in previous genetic studies. Second, the role for three of the variants may plausibly be integrated with current concepts of migraine pathophysiology. rs1835740 modulates glutamate homeostasis [57] and rs11172113 (*LRP1*) [62] may also impact glutamate pathways through interaction of LRP1 receptors with NMDA glutamate receptors. Further, the association at rs10166942 (*TRPM8*) for the first time implicates a gene in a pain related pathway in migraine. Third, the available results are compatible with shared pathophysiology among common forms of migraine, since associations for rs2651899 (*PRDM16*), rs10166942 (*TRPM8*), and rs11172113 (*LRP1*) did not differ for MA and MO [62]. Fourth, the results also suggest differences between population-based and clinic-based forms of migraine. While rs2651899 (*PRDM16*), rs10166942 (*TRPM8*), and rs11172113 (*LRP1*) appear most important for common migraine (i.e., at the population level) [62], rs1835740 may play the major role in more severe forms of migraine (i.e., clinic-based populations) [57]. This agrees with the fact that virtually all migraineurs in the IHGC study suffered from MA [57], which is often considered more severe than MO. However, the association of rs1835740 with MA may also underline the importance of glutamate for the aura phenotype among migraineurs. Finally, the effect estimates indicate that each of the variants only confers a small to moderate change in risk for migraine (Table 1). This agrees with migraine being a heterogeneous disorder, where many gene variants work together to produce the migraine phenotype.

Outlook

Despite enormous efforts, the published GWASs can “only” claim to have identified four genetic variants to be associated with migraine. This number appears small. However, it is imperative to keep in mind the broad clinical phenotype of migraine, which complicates the identification of genetic variants. In addition, we must consider that the implicated genes are both new, replicated in independent samples, and at the same time plausibly involved in migraine pathophysiology.

Additional approaches in studying migraine genetics include focusing on endophenotypes of migraine and investigating larger numbers of migraineurs and non-migraineurs. Trait component analysis and latent class analysis have been used previously [44, 55, 56], and may also be promising approaches for future GWASs. Further, large international collaborations are ongoing aiming at combining genome-wide genetic data on migraine to increase the power to identify additional genetic variants in migraine. Additional variants will likely be identified. Finally, we need to understand that GWASs “only” investigate hundreds of thousands of common and known gene variants. However, there is ample space in the genome that is not investigated. Hence, variants identified may only be markers for migraine, which are in linkage disequilibrium with the causative variants. Further, there may be areas in the genome that are not well represented on the GWAS platforms, but which may plausibly harbor variants involved in migraine pathophysiology. In both scenarios it may be useful to utilize DNA sequencing methods to investigate certain genes or chromosomal regions to obtain a complete map of all variants, including rare variants, in these regions. This may facilitate the discovery of the single best marker in a region or even identify new variants. Considering the large number of DNA base pairs to be sequenced as well as the large number of individuals that may have to investigated, high-throughput sequencing methods may have to be used. However, the limitations associated with each of the available platforms need to be considered and the technology used must be tailored to the specific projects.

Conflict of interest Dr. Schürks has received an investigator-initiated research grant from the Migraine Research Foundation. He has received honoraria from L.E.K. Consulting for telephone surveys and from the American Academy of Neurology for educational material. He is a full-time employee of Bayer HealthCare Germany Pharmaceuticals since August 2011.

Open Access This article is distributed under the terms of the Creative Commons Attribution License which permits any use, distribution and reproduction in any medium, provided the original author(s) and source are credited.

References

1. Haut SR, Bigal ME, Lipton RB (2006) Chronic disorders with episodic manifestations: focus on epilepsy and migraine. Lancet Neurol 5(2):148–157
2. Silberstein SD (2004) Migraine. Lancet 363(9406):381–391
3. Headache Classification Subcommittee of the International Headache Society (2004) The International Classification of Headache Disorders: 2nd edn. Cephalalgia 24(Suppl 1):9–160
4. Bille B (1997) A 40-year follow-up of school children with migraine. Cephalalgia 17(4):488–491
5. Mulder EJ, van Baal C, Gaist D, Kallela M, Kaprio J, Svensson DA, Nyholt DR, Martin NG, MacGregor AJ, Cherkas LF, Boomsma DI, Palotie A (2003) Genetic and environmental influences on migraine: a twin study across six countries. Twin Res 6(5):422–431
6. Svensson DA, Larsson B, Waldenlind E, Pedersen NL (2003) Shared rearing environment in migraine: results from twins reared apart and twins reared together. Headache 43(3):235–244

7. Ligthart L, Boomsma DI, Martin NG, Stubbe JH, Nyholt DR (2006) Migraine with aura and migraine without aura are not distinct entities: further evidence from a large Dutch population study. Twin Res Hum Genet 9(1):54–63
8. Nyholt DR, Gillespie NG, Heath AC, Merikangas KR, Duffy DL, Martin NG (2004) Latent class and genetic analysis does not support migraine with aura and migraine without aura as separate entities. Genet Epidemiol 26(3):231–244
9. Akerman S, Goadsby PJ (2007) Dopamine and migraine: biology and clinical implications. Cephalalgia 27(11):1308–1314
10. Lipton RB, Bigal ME (2005) The epidemiology of migraine. Am J Med 118(1):3S–10S
11. Brandes JL (2006) The influence of estrogen on migraine: a systematic review. JAMA 295(15):1824–1830
12. Schürks M, Rist PM, Kurth T (2010) Sex hormone receptor gene polymorphisms and migraine: a systematic review and meta-analysis. Cephalalgia 30(11):1306–1328
13. De Fusco M, Marconi R, Silvestri L, Atorino L, Rampoldi L, Morgante L, Ballabio A, Aridon P, Casari G (2003) Haploinsufficiency of ATP1A2 encoding the Na+/K+ pump alpha2 subunit associated with familial hemiplegic migraine type 2. Nat Genet 33(2):192–196
14. Dichgans M, Freilinger T, Eckstein G, Babini E, Lorenz-Depiereux B, Biskup S, Ferrari MD, Herzog J, van den Maagdenberg AM, Pusch M, Strom TM (2005) Mutation in the neuronal voltage-gated sodium channel SCN1A in familial hemiplegic migraine. Lancet 366(9483):371–377
15. Ophoff RA, Terwindt GM, Vergouwe MN, van Eijk R, Oefner PJ, Hoffman SM, Lamerdin JE, Mohrenweiser HW, Bulman DE, Ferrari M, Haan J, Lindhout D, van Ommen GJ, Hofker MH, Ferrari MD, Frants RR (1996) Familial hemiplegic migraine and episodic ataxia type-2 are caused by mutations in the Ca2+ channel gene CACNL1A4. Cell 87(3):543–552
16. de Vries B, Frants RR, Ferrari MD, van den Maagdenberg AM (2009) Molecular genetics of migraine. Hum Genet 126(1):115–132
17. Maher BH, Griffiths LR (2011) Identification of molecular genetic factors that influence migraine. Mol Genet Genomics 285(6):433–446
18. Nyholt DR, LaForge KS, Kallela M, Alakurtti K, Anttila V, Farkkila M, Hamalainen E, Kaprio J, Kaunisto MA, Heath AC, Montgomery GW, Gobel H, Todt U, Ferrari MD, Launer LJ, Frants RR, Terwindt GM, de Vries B, Verschuren WM, Brand J, Freilinger T, Pfaffenrath V, Straube A, Ballinger DG, Zhan Y, Daly MJ, Cox DR, Dichgans M, van den Maagdenberg AM, Kubisch C, Martin NG, Wessman M, Peltonen L, Palotie A (2008) A high-density association screen of 155 ion transport genes for involvement with common migraine. Hum Mol Genet 17(21):3318–3331
19. Pietrobon D, Striessnig J (2003) Neurobiology of migraine. Nat Rev Neurosci 4(5):386–398
20. Hamel E (2007) Serotonin and migraine: biology and clinical implications. Cephalalgia 27(11):1293–1300
21. Ramamoorthy S, Bauman AL, Moore KR, Han H, Yang-Feng T, Chang AS, Ganapathy V, Blakely RD (1993) Antidepressant- and cocaine-sensitive human serotonin transporter: molecular cloning, expression, and chromosomal localization. Proc Natl Acad Sci USA 90(6):2542–2546
22. Heils A, Teufel A, Petri S, Seemann M, Bengel D, Balling U, Riederer P, Lesch KP (1995) Functional promoter and polyadenylation site mapping of the human serotonin (5-HT) transporter gene. J Neural Transm 102(3):247–254
23. Heils A, Teufel A, Petri S, Stober G, Riederer P, Bengel D, Lesch KP (1996) Allelic variation of human serotonin transporter gene expression. J Neurochem 66(6):2621–2624
24. Lesch KP, Bengel D, Heils A, Sabol SZ, Greenberg BD, Petri S, Benjamin J, Muller CR, Hamer DH, Murphy DL (1996) Association of anxiety-related traits with a polymorphism in the serotonin transporter gene regulatory region. Science 274(5292): 1527–1531
25. Schürks M, Rist PM, Kurth T (2010) 5-HTTLPR polymorphism in the serotonin transporter gene and migraine: a systematic review and meta-analysis. Cephalalgia 30(11):1296–1305
26. Sicuteri F (1977) Dopamine, the second putative protagonist in headache. Headache 17(3):129–131
27. Wood PB (2008) Role of central dopamine in pain and analgesia. Expert Rev Neurother 8(5):781–797
28. Moskowitz MA (2007) Pathophysiology of headache—past and present. Headache 47(Suppl 1):S58–S63
29. Covelli V, Munno I, Pellegrino NM, Altamura M, Decandia P, Marcuccio C, Di Venere A, Jirillo E (1991) Are TNF-alpha and IL-1 beta relevant in the pathogenesis of migraine without aura? Acta Neurol (Napoli) 13:205–211
30. Covelli V, Munno I, Pellegrino NM, Di Venere A, Jirillo E, Buscaino GA (1990) Exaggerated spontaneous release of tumor necrosis factor-alpha/cachectin in patients with migraine without aura. Acta Neurol (Napoli) 12(4):257–263
31. Perini F, D'Andrea G, Galloni E, Pignatelli F, Billo G, Alba S, Bussone G, Toso V (2005) Plasma cytokine levels in migraineurs and controls. Headache 45(7):926–931
32. Mueller L, Gupta AK, Stein TP (2001) Deficiency of tumor necrosis factor alpha in a subclass of menstrual migraineurs. Headache 41(2):129–137
33. Durham PL (2006) Calcitonin gene-related peptide (CGRP) and migraine. Headache 46(Suppl 1):S3–S8
34. Wilson AG, Symons JA, McDowell TL, McDevitt HO, Duff GW (1997) Effects of a polymorphism in the human tumor necrosis factor alpha promoter on transcriptional activation. Proc Natl Acad Sci USA 94(7):3195–3199
35. Bouma G, Crusius JB, Oudkerk Pool M, Kolkman JJ, von Blomberg BM, Kostense PJ, Giphart MJ, Schreuder GM, Meuwissen SG, Pena AS (1996) Secretion of tumour necrosis factor alpha and lymphotoxin alpha in relation to polymorphisms in the TNF genes and HLA-DR alleles. Relevance for inflammatory bowel disease. Scand J Immunol 43(4):456–463
36. Schürks M, Rist PM, Zee RYL, Chasman DI, Kurth T (2011) Tumor necrosis factor gene polymorphisms and migraine: a systematic review and meta-analysis. Cephalalgia 31:1381–1404
37. Charles A (2009) Advances in the basic and clinical science of migraine. Ann Neurol 65(5):491–498
38. Schürks M, Rist PM, Bigal ME, Buring JE, Lipton RB, Kurth T (2009) Migraine and cardiovascular disease: a systematic review and meta-analysis. BMJ 339:b3914
39. Tietjen GE, Al-Qasmi MM, Shukairy MS (2002) Livedo reticularis and migraine: a marker for stroke risk? Headache 42(5): 352–355
40. Tietjen GE, Al-Qasmi MM, Gunda P, Herial NA (2006) Sneddon's syndrome: another migraine-stroke association? Cephalalgia 26(3):225–232
41. Vanmolkot FH, van Bortel LM, de Hoon JN (2007) Altered arterial function in migraine of recent onset. Neurology 68(19):1563–1570
42. Tietjen EG (2007) Migraine and ischaemic heart disease and stroke: potential mechanisms and treatment implications. Cephalalgia 27(8):981–987
43. Schürks M, Rist PM, Kurth T (2010) MTHFR 677C>T and ACE D/I polymorphisms in migraine: a systematic review and meta-analysis. Headache 50:588–599
44. Nyholt DR, Morley KI, Ferreira MA, Medland SE, Boomsma DI, Heath AC, Merikangas KR, Montgomery GW, Martin NG (2005)

Genomewide significant linkage to migrainous headache on chromosome 5q21. Am J Hum Genet 77(3):500–512

45. Carlsson A, Forsgren L, Nylander PO, Hellman U, Forsman-Semb K, Holmgren G, Holmberg D, Holmberg M (2002) Identification of a susceptibility locus for migraine with and without aura on 6p12.2–p21.1. Neurology 59(11):1804–1807
46. Nyholt DR, Lea RA, Goadsby PJ, Brimage PJ, Griffiths LR (1998) Familial typical migraine: linkage to chromosome 19p13 and evidence for genetic heterogeneity. Neurology 50(5):1428–1432
47. Lea RA, Shepherd AG, Curtain RP, Nyholt DR, Quinlan S, Brimage PJ, Griffiths LR (2002) A typical migraine susceptibility region localizes to chromosome 1q31. Neurogenetics 4(1):17–22
48. Wessman M, Kallela M, Kaunisto MA, Marttila P, Sobel E, Hartiala J, Oswell G, Leal SM, Papp JC, Hamalainen E, Broas P, Joslyn G, Hovatta I, Hiekkalinna T, Kaprio J, Ott J, Cantor RM, Zwart JA, Ilmavirta M, Havanka H, Farkkila M, Peltonen L, Palotie A (2002) A susceptibility locus for migraine with aura, on chromosome 4q24. Am J Hum Genet 70(3):652–662
49. Deprez L, Peeters K, van Paesschen W, Claeys KG, Claes LR, Suls A, Audenaert D, van Dyck T, Goossens D, Del-Favero J, de Jonghe P (2007) Familial occipitotemporal lobe epilepsy and migraine with visual aura. Linkage to chromosome 9q. Neurology 68(23):1995–2002
50. Cader ZM, Noble-Topham S, Dyment DA, Cherny SS, Brown JD, Rice GP, Ebers GC (2003) Significant linkage to migraine with aura on chromosome 11q24. Hum Mol Genet 12(19):2511–2517
51. Russo L, Mariotti P, Sangiorgi E, Giordano T, Ricci I, Lupi F, Chiera R, Guzzetta F, Neri G, Gurrieri F (2005) A new susceptibility locus for migraine with aura in the 15q11–q13 genomic region containing three GABA-A receptor genes. Am J Hum Genet 76(2):327–333
52. Soragna D, Vettori A, Carraro G, Marchioni E, Vazza G, Bellini S, Tupler R, Savoldi F, Mostacciuolo ML (2003) A locus for migraine without aura maps on chromosome 14q21.2–q22.3. Am J Hum Genet 72(1):161–167
53. Bjornsson A, Gudmundsson G, Gudfinnsson E, Hrafnsdottir M, Benedikz J, Skuladottir S, Kristjansson K, Frigge ML, Kong A, Stefansson K, Gulcher JR (2003) Localization of a gene for migraine without aura to chromosome 4q21. Am J Hum Genet 73(5):986–993
54. Wessman M, Terwindt GM, Kaunisto MA, Palotie A, Ophoff RA (2007) Migraine: a complex genetic disorder. Lancet Neurol 6(6):521–532
55. Anttila V, Kallela M, Oswell G, Kaunisto MA, Nyholt DR, Hamalainen E, Havanka H, Ilmavirta M, Terwilliger J, Sobel E, Peltonen L, Kaprio J, Farkkila M, Wessman M, Palotie A (2006) Trait components provide tools to dissect the genetic susceptibility of migraine. Am J Hum Genet 79(1):85–99
56. Anttila V, Nyholt DR, Kallela M, Artto V, Vepsalainen S, Jakkula E, Wennerstrom A, Tikka-Kleemola P, Kaunisto MA, Hamalainen E, Widen E, Terwilliger J, Merikangas K, Montgomery GW, Martin NG, Daly M, Kaprio J, Peltonen L, Farkkila M, Wessman M, Palotie A (2008) Consistently replicating locus linked to migraine on 10q22–q23. Am J Hum Genet 82(5):1051–1063
57. Anttila V, Stefansson H, Kallela M, Todt U, Terwindt GM, Calafato MS, Nyholt DR, Dimas AS, Freilinger T, Muller-Myhsok B, Artto V, Inouye M, Alakurtti K, Kaunisto MA, Hamalainen E, de Vries B, Stam AH, Weller CM, Heinze A, Heinze-Kuhn K, Goebel I, Borck G, Gobel H, Steinberg S, Wolf C, Bjornsson A, Gudmundsson G, Kirchmann M, Hauge A, Werge T, Schoenen J, Eriksson JG, Hagen K, Stovner L, Wichmann HE, Meitinger T, Alexander M, Moebus S, Schreiber S, Aulchenko YS, Breteler MM, Uitterlinden AG, Hofman A, van Duijn CM, Tikka-Kleemola P, Vepsalainen S, Lucae S, Tozzi F, Muglia P, Barrett J, Kaprio J, Farkkila M, Peltonen L, Stefansson K, Zwart JA, Ferrari MD, Olesen J, Daly M, Wessman M, van den Maagdenberg AM, Dichgans M, Kubisch C, Dermitzakis ET, Frants RR, Palotie A, International Headache Genetics C (2010) Genome-wide association study of migraine implicates a common susceptibility variant on 8q22.1. Nat Genet 42(10):869–873
58. Andreou AP, Goadsby PJ (2009) Therapeutic potential of novel glutamate receptor antagonists in migraine. Expert Opin Investig Drugs 18(6):789–803
59. Sanchez-del-Rio M, Reuter U (2004) Migraine aura: new information on underlying mechanisms. Curr Opin Neurol 17(3):289–293
60. Hadjikhani N, Sanchez Del Rio M, Wu O, Schwartz D, Bakker D, Fischl B, Kwong KK, Cutrer FM, Rosen BR, Tootell RB, Sorensen AG, Moskowitz MA (2001) Mechanisms of migraine aura revealed by functional MRI in human visual cortex. Proc Natl Acad Sci USA 98(8):4687–4692
61. Ligthart L, de Vries B, Smith AV, Ikram MA, Amin N, Hottenga JJ, Koelewijn SC, Kattenberg VM, de Moor MH, Janssens AC, Aulchenko YS, Oostra BA, de Geus EJ, Smit JH, Zitman FG, Uitterlinden AG, Hofman A, Willemsen G, Nyholt DR, Montgomery GW, Terwindt GM, Gudnason V, Penninx BW, Breteler M, Ferrari MD, Launer LJ, van Duijn CM, van den Maagdenberg AM, Boomsma DI (2011) Meta-analysis of genome-wide association for migraine in six population-based European cohorts. Eur J Hum Genet 19(8):901–907
62. Chasman DI, Schürks M, Anttila V, de Vries B, Schminke U, Launer LJ, Terwindt GM, van den Maagdenberg AM, Fendrich K, Volzke H, Ernst F, Griffiths LR, Buring JE, Kallela M, Freilinger T, Kubisch C, Ridker PM, Palotie A, Ferrari MD, Hoffmann W, Zee RY, Kurth T (2011) Genome-wide association study reveals three susceptibility loci for common migraine in the general population. Nat Genet 43(7):695–698
63. Proudfoot CJ, Garry EM, Cottrell DF, Rosie R, Anderson H, Robertson DC, Fleetwood-Walker SM, Mitchell R (2006) Analgesia mediated by the TRPM8 cold receptor in chronic neuropathic pain. Curr Biol 16(16):1591–1605
64. Peier AM, Moqrich A, Hergarden AC, Reeve AJ, Andersson DA, Story GM, Earley TJ, Dragoni I, McIntyre P, Bevan S, Patapoutian A (2002) A TRP channel that senses cold stimuli and menthol. Cell 108(5):705–715
65. Dray A (2008) Neuropathic pain: emerging treatments. Br J Anaesth 101(1):48–58
66. Biondi DM (2006) Is migraine a neuropathic pain syndrome? Curr Pain Headache Rep 10(3):167–178
67. Lillis AP, van Duyn LB, Murphy-Ullrich JE, Strickland DK (2008) LDL receptor-related protein 1: unique tissue-specific functions revealed by selective gene knockout studies. Physiol Rev 88(3):887–918
68. Secker-Walker LM, Mehta A, Bain B (1995) Abnormalities of 3q21 and 3q26 in myeloid malignancy: a United Kingdom Cancer Cytogenetic Group study. Br J Haematol 91:490–501
69. Seale P, Kajimura S, Spiegelman BM (2009) Transcriptional control of brown adipocyte development and physiological function—of mice and men. Genes Dev 23(7):788–797

View publication stats

Genomic epidemiology of syphilis in England: a population-based study

Mathew A Beale, Louise Thorn, Michelle J Cole, Rachel Pitt, Hannah Charles, Michael Ewens, Patrick French, Malcolm Guiver, Emma E Page, Erasmus Smit, Jaime H Vera, Katy Sinka, Gwenda Hughes, Michael Marks, Helen Fifer, Nicholas R Thomson**

Summary

Background Syphilis is a sexually transmitted bacterial infection caused by *Treponema pallidum* subspecies *pallidum*. Since 2012, syphilis rates have risen dramatically in many high-income countries, including England. Although this increase in syphilis prevalence is known to be associated with high-risk sexual activity in gay, bisexual, and other men who have sex with men (GBMSM), cases are rising in heterosexual men and women. The transmission dynamics within and between sexual networks of GBMSM and heterosexual people are not well understood. We aimed to investigate if whole genome sequencing could be used to supplement or enhance epidemiological insights around syphilis transmission.

Methods We linked national patient demographic, geospatial, and behavioural metadata to whole *T pallidum* genome sequences previously generated from patient samples collected from across England between Jan 1, 2012, and Oct 31, 2018, and performed detailed phylogenomic analyses.

Findings Of 497 English samples submitted for sequencing, we recovered 240 genomes (198 from the UK Health Security Agency reference laboratory and 42 from other laboratories). Three duplicate samples (same patient and collection date) were included in the main phylogenies, but removed from further analyses of English populations, leaving 237 genomes. 220 (92·8%) of 237 samples were from men, nine (3·8%) were from women, and eight (3·4%) were of unknown gender. Samples were mostly from London (n=118 [49·8%]), followed by southeast England (n=29 [12·2%]), northeast England (n=24 [10·1%]), and southwest England (n=15 [6·3%]). 180 (76·0%) of 237 genomes came from GBMSM, compared with 25 (10·5%) from those identifying as men who have sex with women, 15 (6·3%) from men with unrecorded sexual orientation, nine (3·8%) from those identifying as women who have sex with men, and eight (3·4%) from people of unknown gender and sexual orientation. Phylogenomic analysis and clustering revealed two dominant *T pallidum* sublineages in England. Sublineage 1 was found throughout England and across all patient groups, whereas sublineage 14 occurred predominantly in GBMSM older than 34 years and was absent from samples sequenced from the north of England. These different spatiotemporal trends, linked to demography or behaviour in the dominant sublineages, suggest they represent different sexual networks. By focusing on different regions of England we were able to distinguish a local heterosexual transmission cluster from a background of transmission in GBMSM.

Interpretation These findings show that, despite extremely close genetic relationships between *T pallidum* genomes globally, genomics can still be used to identify putative transmission clusters for epidemiological follow-up. This could be of value for deconvoluting putative outbreaks and for informing public health interventions.

Funding Wellcome funding to the Sanger Institute, UK Research and Innovation, National Institute for Health and Care Research, European and Developing Countries Clinical Trials Partnership, and UK Health Security Agency.

Crown Copyright © 2023 Published by Elsevier Ltd. This is an Open Access article under the CC BY 4.0 license.

Lancet Microbe 2023; 4: e770–80

Published **Online**
September 15, 2023

https://doi.org/10.1016/S2666-5247(23)00154-4

*Contributed equally

Parasites and Microbes Programme, Wellcome Sanger Institute, Hinxton, UK (M A Beale PhD, Prof N R Thomson PhD); **Blood Safety, Hepatitis, STI & HIV Division** (L Thorn MPH, H Charles MSc, K Sinka PhD, H Fifer FRCPath) **and HCAI, Fungal, AMR, AMU and Sepsis Division,** (M J Cole FRCPath, R Pitt MSc), **UK Health Security Agency, London, UK; Brotherton Wing Clinic, Brotherton Wing, Leeds General Infirmary, Leeds, UK** (M Ewens MRCP); **The Mortimer Market Centre, Central and North West London NHS Trust, London, UK** (P French FRCP); **Laboratory Network, Manchester, UK Health Security Agency, Manchester Royal Infirmary, Manchester, UK** (M Guiver FRCPath); **Virology Department, Old Medical School, Leeds Teaching Hospitals Trust, Leeds, UK** (E E Page MD); **Clinical Microbiology Department, Queen Elizabeth Hospital, Birmingham, UK** (E Smit FRCPath); **Institute of Environmental Science and Research, Wellington, New Zealand** (E Smit); **Department of Global Health and Infection, Brighton and Sussex Medical School, University of Sussex, Brighton, UK** (Prof J H Vera PhD); **Department of Infectious Disease Epidemiology** (Prof G Hughes PhD), **Faculty of Infectious and Tropical Diseases** (M Marks PhD, Prof N R Thomson), **London School of Hygiene & Tropical Medicine, London, UK; Hospital for Tropical Diseases, University College London Hospitals NHS Foundation Trust, London, UK** (M Marks); **Division of Infection and Immunity, University**

Introduction

Syphilis is a sexually transmitted infection (STI) caused by the bacterium *Treponema pallidum* subspecies *pallidum* (hereafter referred to as *T pallidum*). Syphilis rates have been rising in many high-income countries since the beginning of the 21st century.[1–4] In England, new diagnoses of early syphilis (primary, secondary, and early latent) rose from 3011 in 2012 (5·6 per 100 000 population) to 8011 in 2019 (14·2 per 100 000 population).[3] This increase has primarily been associated with gay, bisexual, and other men who have sex with men (GBMSM) engaging in high-risk sexual behaviours.[2,3,5] However, cases of syphilis in heterosexual men and women have also risen, raising concerns about infection during pregnancy and risks of vertical transmission leading to congenital syphilis.[6,7] Between 2016 and 2019, annual syphilis diagnoses increased by 53% (from 660 to 1012) in men who identify as heterosexual and have sex with women (MSW) and 108% (from 294 to 614) in women who identify as heterosexual and have sex with men (WSM). 24 cases of congenital syphilis were identified in England between 2015 and 2020, 15 of which were in children born to mothers who tested negative at first trimester

College London, London, UK
(M Marks)

Correspondence to:
Dr Mathew Beale, Parasites & Microbes Programme, Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridgeshire, CB10 1SA, UK
mathew.beale@sanger.ac.uk

or

Dr Helen Fifer, Blood Safety, Hepatitis, STI & HIV Division, UK Health Security Agency, London NW9 5EQ, UK
helen.fifer@ukhsa.gov.uk

Research in context

Evidence before this study

Detailed phylogenomic analyses investigating the epidemiology and transmission dynamics of *Treponema pallidum* are challenging due to low bacterial loads in clinical specimens, and difficulty in culturing the bacteria. We searched PubMed from Jan 1, 1998, to Aug 9, 2022, using the search terms "Syphilis" or "*Treponema pallidum*" and "genomic" or "genome(s)" or "sequencing". We found 23 studies describing whole genome sequencing of the *T pallidum* subspecies *pallidum*, of which two used whole genome phylogenies to investigate sexual network epidemiology, with one large study of sexual networks done primarily in Victoria, Australia, which characterised two major circulating sublineages in that setting and putative sexual transmission networks with distinct sexual behavioural characteristics and potential bridging between networks.

Added value of this study

In this study, we linked national surveillance data to *T pallidum* genomes, and characterised the transmission dynamics of syphilis using samples from across the whole of England. Integration of national-level sociodemographic, spatiotemporal, and genomic data allowed the delineation of putative sexual networks at both the national and region levels, and revealed patterns not previously detected using epidemiological or genomic data alone.

Implications of all the available evidence

Our findings are consistent with findings in Australia that indicate that genomics can identify putative sociodemographic transmission clusters. However, in that study, genomic clusters included samples separated by multiple single nucleotide polymorphisms, which could represent several years of evolution. Our study explored the value of linking identical genomes, and highlights that, despite technical constraints, whole genome sequencing can be used to enable outbreak exclusion and identify putative local transmission clusters for epidemiological follow-up.

antenatal screening,[8] indicating they had acquired syphilis later in pregnancy. Some cases were identified in regions across England with increases in syphilis among women and GBMSM, suggesting that overlapping sexual networks might have facilitated wider dissemination.[9]

Although epidemiological surveillance provides insights into the rise in syphilis rates, this is not always sufficient. For example, a group of spatiotemporally clustered cases could represent a single outbreak and chain of transmission, but could also be the result of separate or unrelated transmission networks. Molecular typing methods[10–13] provide one possible way to supplement epidemiological observations by identifying genetically related *T pallidum* strains. However, these methods might not accurately reflect recent evolutionary relationships between strains,[14–16] but instead cluster groups of bacteria that shared a common ancestor decades ago, meaning it would be impossible to accurately delineate strain clusters relevant to epidemiologically useful timelines (usually months to years).

Whole genome sequencing (WGS) has shown there are two co-circulating *T pallidum* lineages globally (Nichols and SS14),[17–19] which can be further divided into 17 sublineages plus singletons.[17] These data showed that *T pallidum* genomes accumulate single nucleotide polymorphisms (SNPs) very slowly, with a median molecular clock rate equivalent to one substitution per genome every 6·9 years (95% highest posterior density 5·9–8·2 years),[17] similar to other studies,[20,21] which means that isolate genomes from strains circulating in the UK can be identical (zero pairwise-SNPs) to those from Canada, Australia, and other countries, and identical genomes were collected an average of 2·5 years apart (range 0–15 years). This genetic homogeneity has been suggested to indicate a global dissemination of *T pallidum* within the past 30 years, driven by a small number of multi-country sublineages,[17] and that genomic approaches might also be of limited value to investigate or resolve epidemiological links between patients with syphilis.

See **Online** for appendix 1

Few studies have combined WGS with patient demographic and sexual behaviour metadata to explore epidemiological trends of *T pallidum*, with studies from Japan[22] and Australia[21] finding discrete genetic clusters associated with GBMSM and heterosexual people. We explored the value of WGS for supplementing existing epidemiological data for understanding transmission at national and regional levels. We combined detailed patient demographic and epidemiological data with WGS of *T pallidum* samples from England to gain insights into the different spatiotemporal and genomic transmission patterns of syphilis affecting GBMSM and heterosexual people.

Methods

Study design and participants

A detailed description of sample collection and patient metadata linkage in this population-based study is provided in appendix 1 (pp 1–2). *T pallidum*-positive genomic DNA samples were retrieved from historical archives (2012–17) held at the UK Health Security Agency (UKHSA, previously Public Health England) STI Reference Laboratory (Colindale, London), and prospectively collected (2016–18) from five laboratories with high syphilis caseloads who do in-house molecular *T pallidum* diagnostic testing (and thus do not usually refer to the UKHSA reference laboratory; Birmingham, Brighton, Leeds, London [Mortimer Market Clinic at Central and North West London NHS Trust], and Manchester).

For samples from UKHSA, patient metadata were obtained by linkage to the national STI surveillance system (Genitourinary Medicine Clinic Activity Dataset [GUMCAD]). For samples prospectively collected from the five non-referring laboratories, patient metadata available from local laboratory information systems was linked to GUMCAD data and integrated into the larger dataset after deduplication (appendix 1 pp 1–2). For comparison between the sequencing dataset and national surveillance rates, we also retrieved summary statistics from GUMCAD data for all patients with syphilis aged 16 years and older in England from 2012 to 2018 (n=50 845).

Ethics and data governance

Ethical approval for all clinical samples was granted by the National Health Service (UK) Health Research Authority and Health and Care Research Wales (UK; 19/HRA/0112) and the London School of Hygiene & Tropical Medicine Observational Research Ethics Committee (16014). Diagnostic samples were identified at UKHSA using internal laboratory information systems. UKHSA has permission to process confidential patient data under regulation 3 (control of patient information) of the UK Health Service Regulations 2002. Information governance advice and ethics approval for this study were granted by the UKHSA Research Ethics and Governance Group. Full details of approvals and pseudonymisation of samples and patient metadata are described in appendix 1 (pp 2–3).

Procedures

WGS of all clinical *T pallidum* samples used in this study has been previously described,[18] and was done directly on the genomic DNA extracts from residual diagnostic samples using the pooled sequence capture method[19] on Illumina HiSeq 4000. Detailed spatiotemporal and sociodemographic metadata retrieved from the national STI surveillance system were linked to the multiple sequence alignments and time-scaled phylogenies generated and validated previously, to maintain consistency (appendix 1 pp 1–5).[17]

Outcomes

The primary outcome was to identify local lineages of *T pallidum* samples and distinguish transmission clusters by location or sexual orientation. A secondary outcome was to investigate rates and distribution of macrolide resistance-conferring alleles among *T pallidum* sublineages.

Statistical analysis

All statistical analyses were done in R, version 4.1.2 (appendix 1 p 5).

Role of the funding source

The funders of the study had no role in study design, data collection, data analysis, data interpretation, writing of the report, or decision to submit.

Results

Between Jan 1, 2012, and Oct 31, 2018, we identified and submitted 497 samples considered suitable for sequencing, from which we recovered 240 genomes (198 from the UKHSA reference laboratory and 42 from other laboratories; appendix 1 p 9). Three duplicate samples (same patient and collection date) were included in the main phylogenies, but removed from further analyses of English populations, leaving 237 genomes. 220 (92·8%) of 237 samples were from men, nine (3·8%) were from women and eight (3·4%) were of unknown gender. 217 samples were grouped into nine official Public Health Regions; for 18 samples the region was unknown; two samples were referred to English laboratories from elsewhere in the UK (table). Most samples were from London (n=118 [49·8%]), followed by southeast England (n=29 [12·2%]), northeast England (n=24 [10·1%]), and southwest England (n=15 [6·3%]). Analysis of collection dates showed that although we had sequences from most regions throughout the study period, the northwest and Yorkshire and Humber were represented only at the beginning (2012 and 2013) and end (2018) of the timeline.

For more on **GUMCAD** see https://www.gov.uk/guidance/gumcad-sti-surveillance-system

180 (75·9%) of 237 genomes came from GBMSM, compared with 25 (10·5%) from MSW, 15 (6·3%) from men with unrecorded sexual orientation, nine (3·8%) from WSM, and eight (3·4%) from people of unknown gender and sexual orientation (table; appendix 1 p 10). Notably, the most highly represented region (London) had a higher proportion of GBMSM (110 [93%]) compared with the next most highly represented regions (21 [72·4%] from the southeast, 15 [62·5%] from the northeast, and 12 [80·0%] from the southwest). Due to a low number of heterosexual individuals (30 of 195) in the UKHSA dataset, HIV status was restricted to GBMSM to prevent deductive disclosure (the prospectively collected samples included no MSW or WSM living with HIV). 65 (27·4%) of 237 samples in the national genome collection were from people living with HIV and these were distributed across seven of nine regions. In London, which had the highest proportion of GBMSM, 48 (40·7%) patients were living with HIV.

To establish how representative our national genome collection was of syphilis cases in England, we compared the distributions of sociodemographic characteristics of cases in the national genome collection with those from GUMCAD (table). Overall, our national genome collection (n=237) represented 0·5% of all syphilis diagnoses among patients aged 16 years and older in England during the period 2012–18 (n=50 845). Compared with diagnoses reported in GUMCAD, the samples used in the WGS project were broadly representative by age

	Whole genome sequencing (n=237)	All syphilis diagnoses (n=50845)
Year		
2012	21 (8·9%)	4856 (9·6%)
2013	19 (8·0%)	5308 (10·4%)
2014	12 (5·1%)	6342 (12·5%)
2015	38 (16·0%)	7351 (14·5%)
2016	71 (30·0%)	8034 (15·8%)
2017	48 (20·3%)	9177 (18·0%)
2018	28 (11·8%)	9777 (19·2%)
Self-identified gender		
Male	220 (92·8%)	42 168 (82·9%)
Female	9 (3·8%)	6467 (12·7%)
Unknown	8 (3·4%)	2210 (4·3%)
Gender orientation		
MSW	25 (10·5%)	8978 (17·7%)
GBMSM	180 (75·9%)	33190 (65·3%)
WSM	9 (3·8%)	6467 (12·7%)
Men with unrecorded sexual orientation	15 (6·3%)	..
Unknown	8 (3·4%)	2210 (4·4%)
Age group		
16–24 years	30 (12·7%)	6194 (12·2%)
25–34 years	70 (29·5%)	16298 (32·1%)
35–44 years	59 (24·9%)	13487 (26·5%)
≥45 years	76 (32·1%)	14671 (28·9%)
Unknown	2 (0·8%)	195 (0·4%)
Region of residence		
East Midlands	5 (2·1%)	2398 (4·7%)
East of England	8 (3·4%)	2495 (4·9%)
London	118 (49·8%)	24326 (47·8%)
Northeast	24 (10·1%)	1497 (2·9%)
Northwest	6 (2·5%)	5079 (9·5%)
Southeast	29 (12·2%)	4776 (9·4%)
Southwest	15 (6·3%)	2082 (4·1%)
West Midlands	7 (3·0%)	3763 (7·4%)
Yorkshire & Humber	5 (2·1%)	2760 (5·4%)
Unknown	18 (7·6%)	1475 (2·9%)
UK (not England)	2 (0·8%)	194 (0·4%)
UK birth status		
Non-UK born	66 (27·8%)	24191 (47·6%)
UK born	138 (58·2%)	26654 (52·4%)
Unknown	33 (13·9%)	..
HIV status (GBMSM only)*		
Negative	121 (51·1%)	21888 (43·1%)
Positive	65 (27·4%)	11302 (22·2%)
Data unavailable	51 (21·5%)	17655 (34·7%)
Syphilis stage (matched to diagnoses only, n=140†)		
Primary	114 (81·4%)	14178 (27·9%)
Secondary	15 (10·7%)	9918 (19·5%)
Early latent	10 (7·1%)	12261 (24·1%)
Late latent	1 (0·7%)	13032 (25·6%)
Cardio	0	635 (1·3%)
Neuro	0	821 (1·6%)

(Table continues in next column)

(Continued from previous column)

Gender categories are based on reported sexual orientation and gender identity, not on behaviours. WGS=whole genome sequencing. GUMCAD=Genitourinary Medicine Clinic Activity Dataset. MSW=men who have sex with women. GBMSM=gay, bisexual, and other men who have sex with men. WSM=women who have sex with men. *All diagnoses: n=33 190; WGS: n=186. †WGS dataset only.

***Table:* Baseline characteristics of WGS and all syphilis diagnoses from GUMCAD in England, 2012–18**

group, region of residence (including London *vs* non-London), and HIV status (GBMSM only). However, a greater proportion of cases in the genome collection were GBMSM (180 [75·9%] of 237 from the WGS project *vs* 33 190 [65·3%] of 50 845 in the general population), with fewer women (nine [3·8%] *vs* 6467 [12·7%]) and MSW (25 [10·5%] *vs* 8978 [17·7%]). The genome collection also had a much higher proportion of primary syphilis cases compared with GUMCAD (114 [81·4%] in the genome collection compared with 14 178 [27·9%] from GUMCAD), largely reflecting the clinical presentation of primary syphilis with ulcers that permit swabbing.

We inferred the presence of macrolide resistance-conferring SNPs in the ribosomal 23S as previously described,[18] and found that 209 (88·2%) of 237 English genome samples carried the A2058G allele, six (2·5%) of 237 had an uncertain or mixed variant call at position 2058, and five (2·1%) of 237 carried the A2059G allele, meaning that only 17 (7·2%) of 237 English *T pallidum* genomes carried a wild type ribosomal 23S gene and were therefore predicted to be sensitive to macrolide antimicrobials.

A whole genome phylogeny was inferred from the 237 English genomes sequenced here, along with 286 global contextual genomes. We clustered all isolates by lineage, sublineage, or into single-linkage SNP clusters.[23] The English genomes were broadly distributed throughout the known *T pallidum* phylogeny (appendix 1 p 11). Referencing previous work by us and others,[17–19] 183 (77·2%) of 237 genomes belonged to the SS14 lineage and 54 (23·8%) belonged to the Nichols lineage. Of the 17 defined global sublineages,[17] eight were present in the UK, along with one singleton (figure 1A–C; appendix 1 p 12). The English samples were dominated by the global sublineages 1 (n=175 [73·8%]) and 14 (n=44 [18·6%]), but two other globally distributed sublineages (five from sublineage 2 and five from sublineage 8) were also detected in the UK (figure 1A, C) as well as two English samples for each of sublineages 3, 6, and 15, and one for sublineage 16.

Linking previously sequenced whole *T pallidum* genomes from English patients to STI surveillance and

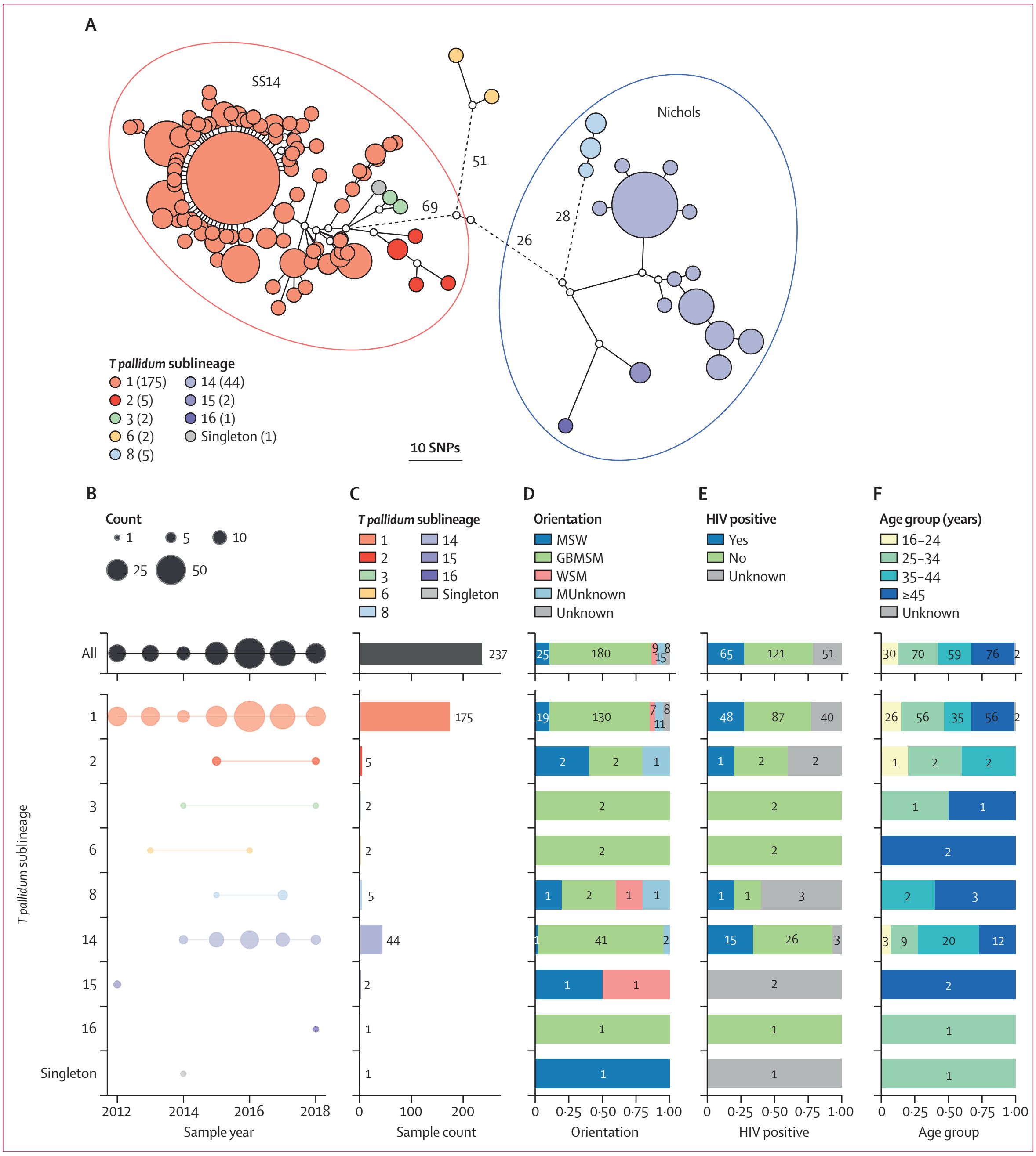


Figure 1: **Population structure of English *Treponema pallidum* genomes according to phylogenetic sublineages and associated patient characteristics**
(A) Minimum spanning tree visualisation of genetic relationships between samples from England. Node size corresponds to the number of identical genome samples in a cluster, and edge length (with number) to the number of substitutions between identical genome sample clusters (where edges were longer than 12 substitutions, these have been shortened; this is indicated by dashed lines). Numbers in parentheses indicate total sample count for each sublineage. Primary lineage (SS14/Nichols) is indicated by encompassing ellipses; sublineage 6 diverges from other *Treponema pallidum* subspecies *pallidum* close to the root, and has previously been classified as Nichols.[18] (B) Samples per collection year per sublineage. (C) Total sample counts per sublineage. Bar plots show proportion of each group. (D) Proportion of each group by sexual orientation. (E) Proportion of each group by HIV status. (F) Proportion of each group by age group (numbers indicate exact sample counts). SNP=single nucleotide polymorphism. MSW=men who have sex with women. GBMSM=gay, bisexual, and other men who have sex with men. WSM=women who have sex with men. MUnknown=men with unrecorded sexual orientation.

laboratory records, we observed that patients infected with the most common sublineage (175 [73·8%] of 237 from sublineage 1) were largely representative of syphilis patients overall, with 130 (74·3%) of 175 classified as GBMSM (figure 1D, appendix 1 p 12) and 91 (52·0%) aged 35 years or older. By contrast, 41 (93·2%) of

44 patients infected with sublineage 14 were GBMSM (1 of 44 MSW, 2 of 44 men with unrecorded sexual orientation), significantly more than would be expected by chance (Fisher's exact p=0·0087). 32 (72·7%) of 44 patients were aged 35 years or older, and 15 (34·1%) were living with HIV. We also found that most patients living with HIV were infected with sublineages 1 or 14, consistent with these lineages being linked to GBMSM networks (figure 1E, appendix 1 p 13). Seven (77·8%) of nine women were infected with *T pallidum* sublineage 1 and no women were infected with sublineage 14 (figure 1D, appendix 1 p 12). We found some rarer sublineage groups contained a greater proportion of heterosexual men and women, with sublineage 8 containing two heterosexual people (one [20·0%] of five WSM, one [20·0%] of five MSW), and sublineage 15 containing two heterosexual people (one [50·0%] of two WSM, one [50·0%] of two MSW) residing in the east of England. Analysis of the genetic relationships indicated that, at least in these two examples, the heterosexual samples were genetically identical to one another but distinct from other samples, falling on terminal nodes of our minimum spanning network (appendix 1 p 12).

To explore sociodemographic patterns of *T pallidum* genome clustering further, we delineated genomes from English patients into 27 distinct clusters of two or more genomes with zero pairwise-SNPs between them (ie, identical at the core genome level; appendix 1 p 14). Given the genetic homogeneity of *T pallidum*, these clusters do not necessarily indicate direct patient-to-patient transmission, but instead provided a means of clustering samples sharing a recent common ancestor. Of the 146 genomes included in these zero-distance genome clusters, 20 were from patients identifying as heterosexual (15 MSW and five WSM). 11 (55·0%) of these 20 patients were part of a cluster containing only other heterosexual individuals. Of the four genome clusters containing WSM, three contained only other heterosexual people and each of these clusters was detected in only a single region. However, the fourth cluster containing a WSM was the largest cluster of identical genomes, comprising 42 samples from sublineage 1, of which 35 patients (83·3%) were GBMSM, compared with three MSW, one WSM, and three with unknown sexual orientation. Therefore, this primarily GBMSM cluster also includes four patients who identify as heterosexual, indicating that there might be some bridging between the populations.

To understand the global context of *T pallidum* from England, we used a Bayesian time-scaled phylogeny of the global dataset (including contextual genomes from 21 other countries, appendix 1 p 15), and examined subtrees for the four most common English sublineages (1, 2, 8, and 14; appendix 1 p 16). Sublineage 1, previously found to be globally disseminated,[17] contained samples from around England and the world; the English samples were polyphyletic and distributed throughout the sublineage 1 phylogeny, with little clustering (appendix 1 p 16), with the exception of a clade predominantly comprising samples from northeast England. Consistent with previous observations, 157 (89·7%) of 175 sublineage 1 strains carried the ribosomal 23S A2058G allele, and were predicted to be resistant to macrolides[17,18] (appendix 1 p 17). As previously described,[17] sublineage 2 comprises two clades, one of which is dominated by North American samples, with the other dominated by samples from China. English samples were found within both clades, with at least three distinct groupings of English strains (appendix 1 p 16), probably indicating multiple recent independent introductions from other countries. By contrast, we found five sublineage 8 samples from England forming a monophyletic subclade along with individual samples from Canada and Australia (appendix 1 p 16). Sublineage 14 represented a major English sublineage, with 44 of 55 sublineage 14 genomes in the global collection coming from England, all of which had the ribosomal 23S A2058G allele. We previously described the contemporaneous appearance of this sublineage in England and Canada in 2013 to 2014,[17] but our time-scaled phylogeny shows two clades within sublineage 14, both of which have median time to most recent common ancestors (1999 and 2006) predating the first detection in our dataset, suggesting this sublineage had been circulating in England for some time (appendix 1 p 16).

We examined the geographical distribution of types in England and found that both the SS14 and Nichols lineages were co-circulating in London and throughout south and central England (figure 2A, E). However, we found only the SS14 lineage in the three most northerly regions (northeast, northwest, Yorkshire and Humber). Examination of *T pallidum* sublineage distributions indicated that sublineage 1 (SS14 lineage) was present in all regions (and represented the only sublineage present in the three northern regions), while sublineage 14 (Nichols lineage) was co-circulating with sublineage 1 in London, the south, and central England but not in northern England (figure 2B, F). There were 35 samples from the three northerly regions, of which 17 (48·6%) were collected after the first detected appearance of sublineage 14 in 2014, coinciding with an increase in national syphilis rates.[24] From 2014 to 2018, the prevalence of sublineage 14 within the national genome collection was 22% (44 of 197). Under an assumption of even sample coverage and sublineage distributions, we could reasonably expect four samples (95% CI 1–7, p=0·022 for zero under a Poisson distribution) to be found in the northern regions, and the absence suggests regional strain distribution is not homogenous.

Apart from the three northern regions and the West Midlands (which contained only sublineages 1 and 14), all other regions contained at least three sublineages (range 3–7). 118 (49·7%) of 237 samples were from

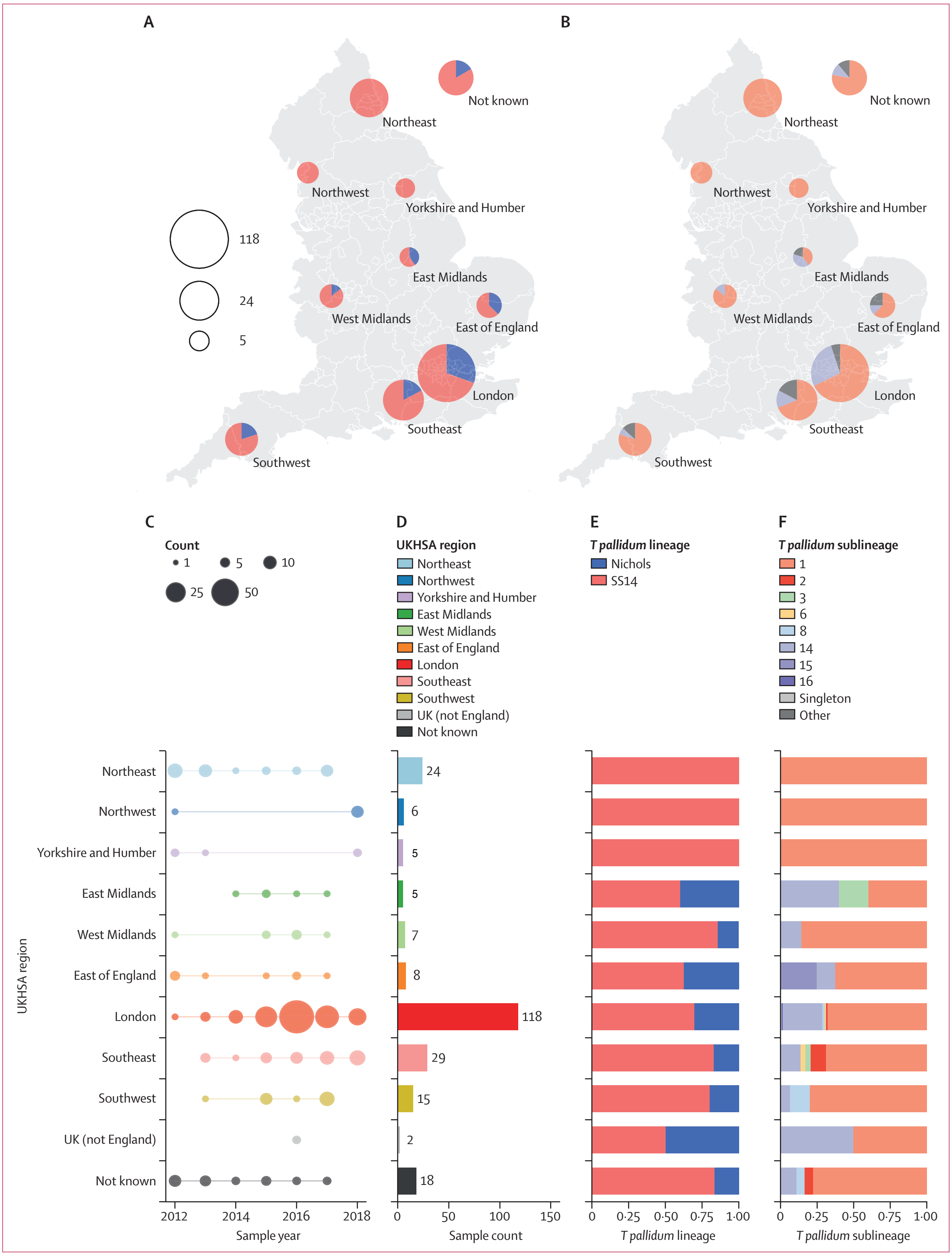


***Figure 2:* Geographical distribution of English genome samples and according to phylogenetic sublineages**
(A) Proportion of samples from each UKHSA region of England by *Treponema pallidum* subspecies *pallidum* lineage. (B) Proportion of samples in each public health region of England for the two most *common T pallidum* sublineages. (C) Distribution of sample collection years. (D) Total sample counts. (E) Proportion of samples from each region by lineage. (F) Proportion of samples from each region by sublineage. UKHSA=UK Health Security Agency.

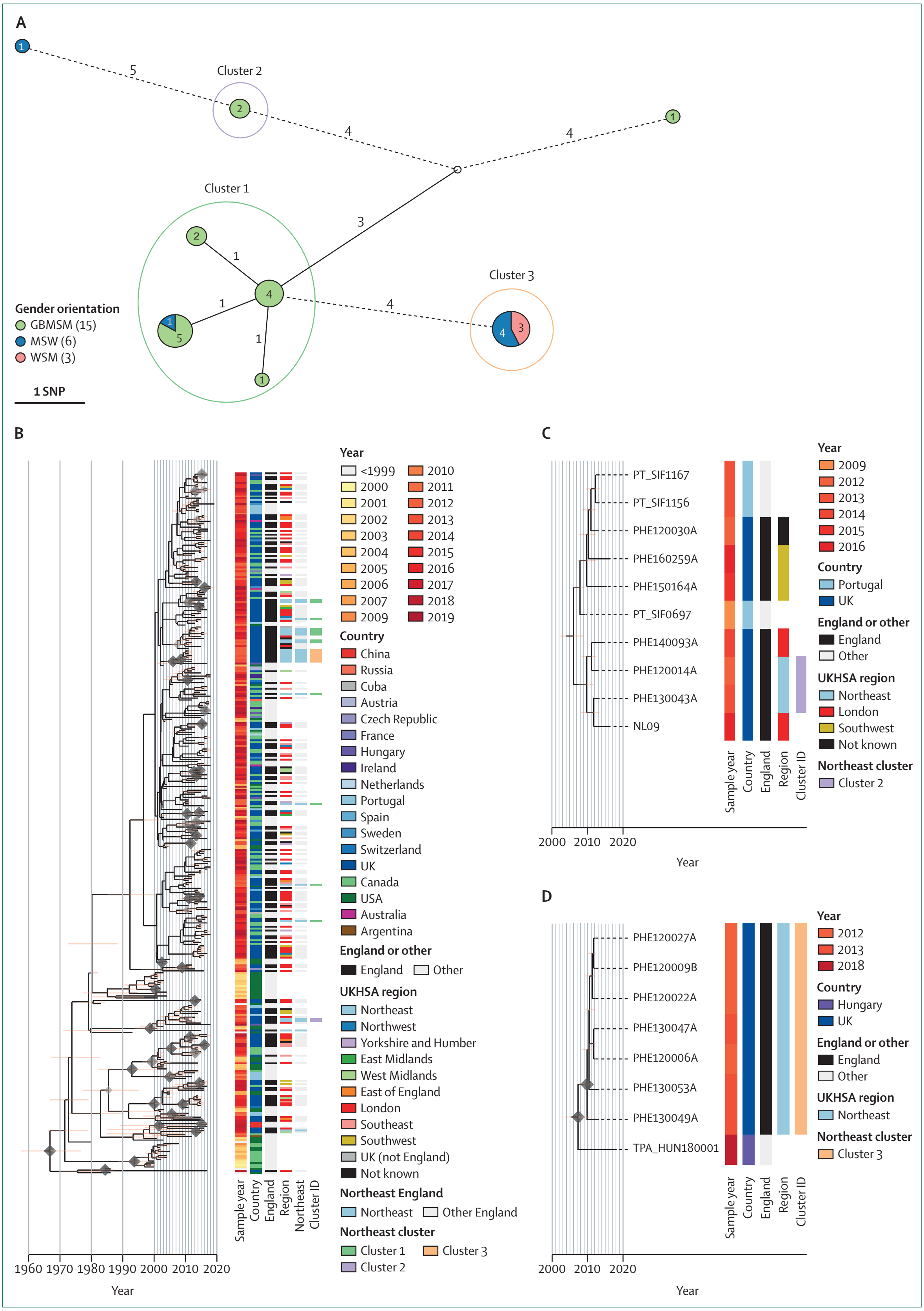


Figure 3: **Spatiotemporal and genomic clustering analysis of samples from northeast England**

(A) Minimum spanning tree visualisation of genomic relationships between samples in northeast England. Node size corresponds to the number of identical samples, and edge length (with number) to the number of substitutions between clusters (where edges were longer than 3 substitutions, these have been shortened, indicated by dashed lines). Nodes are coloured by proportion of patient gender orientation). Clusters were defined by connections to another sample within two pairwise-single nucleotide polymorphisms. Clusters 1 and 2 appear to be dominated by GBMSM populations, while Cluster 3 contains only patients identifying as heterosexual. (B) Time-scaled sublineage 1 tree of global samples indicates that the GBMSM-associated cluster 1 is a globally distributed cluster, and that the northeast samples are polyphyletic. (C) Time-scaled subtree of global samples sharing a common ancestor with cluster 2 indicates a close relationship with two samples from London, and more distantly with those from southwest England and Portugal. (D) Time-scaled subtree of global samples sharing a common ancestor with the heterosexual-associated cluster 3 suggests the northeast England samples are closely related to each other, but not to any others, with the closest related strain found in Hungary. This could imply a historic importation from another country, followed by local circulation. MSW=men who have sex with women. GBMSM=gay, bisexual and other men who have sex with men. WSM=women who have sex with men. UKHSA=UK Health Security Agency.

London, and we detected six sublineages (1, 2, 6, 8, 14, and 16) and one sublineage previously labelled as a singleton. As elsewhere, London was dominated by sublineages 1 (n=80, 67·8%) and sublineage 14 (n=32, 27·1%). The rare sublineages 15 (n=2) and 16 (n=1) were each found in a single region (sublineage 15 in the east of England and sublineage 16 in London), but all other sublineages were found in multiple regions (figure 2).

Our geospatial analysis showed that all samples from the three most northerly regions of England were from SS14 sublineage 1 (figure 2), contrasting with greater diversity elsewhere in England. To explore this in more detail, we focused on the 24 samples collected from the northeast of England between 2012 and 2017. Although all samples belonged to sublineage 1, grouping northeast samples using a two pairwise-SNP threshold identified three distinct northeast clusters within sublineage 1 (figure 3A), and correlation with the global phylogeny (figure 3B; appendix 1 p 16) indicated different distributions (figure 3B). Cluster 1 comprised samples from 13 people that were collected between 2012 and 2017, 12 of which were samples from GBMSM. Within this cluster, samples could be further subdivided into three subgroups of eleven samples linked by identical core genomes (zero pairwise SNPs; figure 3A). Phylogenetic analysis of all English and global samples from sublineage 1 showed that, although posterior support for internal nodes was low, the northeast samples appeared to be polyphyletic (figure 3B) and interspersed with samples (including additional identical genomes) from the rest of England and around the world. Therefore, it is unlikely that the northeast samples from cluster 1 represent a direct chain of transmission or local outbreak, but rather that we have sampled from a broader transmission network spanning national and global boundaries.

Cluster 2 comprised two samples collected in 2012 and 2013 from GBMSM with identical core genomes and, similarly to cluster 1, formed a clade with eight other samples, of which five were from elsewhere in England (four of five GBMSM and one of five unknown) and three were from Portugal[25] (figure 3C). By contrast, cluster 3 comprised seven samples, all of which came from the northeast in 2012 and 2013, and had identical core genomes (zero pairwise-SNPs). These samples formed a monophyletic clade, with the nearest related other sample in the global dataset separated by two SNPs and isolated from a bisexual man in Hungary in 2018 (figure 3D).[17] All seven cluster 3 patients self-identified as heterosexual (MSW or WSM). Given the close spatiotemporal and genomic relationships between cluster 3 samples, and contextualised by a background of greater diversity over the 2012–17 timespan of all other samples collected from northeast England, cluster 3 probably represents a localised outbreak in a heterosexual network. This observation, made solely on the basis of the available genomic and sociodemographic data, is consistent with reports of a syphilis outbreak in heterosexual people in the northeast.[26] It is likely that some of our samples are derived from this event.

Discussion

In this study, we linked patient demographic, spatiotemporal, and behavioural metadata to previously generated *T pallidum* genomes from 237 patients diagnosed with syphilis in England between 2012 and 2018. Our analysis shows a variety of English sublineages, dominated by global sublineages 1 and 14,[17] both of which are predicted to be resistant to macrolides, consistent with the high percentage of macrolide resistant samples in the UK. The English sublineages 1 and 14 displayed different patient sociodemographic and spatiotemporal profiles, with sublineage 1 patients showing a greater diversity of gender, sexual orientation, HIV status, and age, while sublineage 14 was primarily found in older GBMSM. Moreover, although sublineage 1 was found in all regions of England, cases attributed to sublineage 14 were mainly taken in London, and not found in the northern regions of England. These contrasting characteristics suggest that the two sublineages describe distinct sexual transmission networks, consistent with a recent WGS study from Australia,[21] which identified broadly similar *T pallidum* population structures co-circulating in Melbourne and the Northern Territory. Both common sublineages (1 and 14) contained people living with and without HIV, and there was no phylogenetic delineation by HIV status, suggesting either that HIV status might not be strongly associated with transmission patterns, or that such patterns are beyond the ability of WGS-based analyses to detect.

We were also able to examine whether the data could be used to define GBMSM and heterosexual transmission networks based on the proportion of individuals identified as GBMSM or heterosexual for each genomic cluster.[21] We observed three instances in which genomes from heterosexual individuals clustered with identical genomes only from other heterosexual people from the same region, consistent with this representing discrete heterosexual transmission networks or clusters. By contrast, we found that many genetic clusters classified as GBMSM-associated under a proportional definition across the whole dataset exhibited spatiotemporal diversity. These differences in spatiotemporal diversity could reflect differences in partner seeking behaviour and partner concurrency between GBMSM and heterosexual people.[27] We also found mixed clusters, in particular a large cluster of 42 samples with identical core genomes, most of which were from GBMSM, four from heterosexual people (one woman, three men) and three with unreported sexual orientation or gender. Samples in this cluster had diverse regional geography and spanned across the 7 years of this study, and this implies widespread dissemination through the population more rapidly than the bacteria acquires variation, and potentially represents multiple local transmission networks all sharing a recent common ancestor. The presence of heterosexual people within these networks indicates possible bridging between transmission network groups.[28]

As in most countries, samples from England were dominated by sublineage 1.[17,21] Although most English sublineage 1 patients were GBMSM, with *T pallidum* genomes occupying positions across the sublineage 1 phylogeny and interspersed with samples from around the world, in the northeast of England we identified a genetically distinct cluster of identical core genomes found exclusively in heterosexual people, consistent with reports of a syphilis outbreak in heterosexual people at that time.[26] Given the previous uncertainty as to whether genomics can have a substantial role in understanding the epidemiology of syphilis due to the genetic homogeneity and low molecular clock rate of *T pallidum*,[17,20,21] our identification of discrete clusters associated with sexual behaviour suggests WGS combined with detailed epidemiological data can resolve some local transmission chains for *T pallidum*. This could offer opportunities to intervene or educate sexual networks, and to determine or exclude outbreak membership.

In other STIs, such as gonorrhoea, SNP cutoffs of either five or ten SNPs have been used to infer transmission chains.[28,29] *Neisseria gonorrhoeae* accumulates SNPs at a rate of eight substitutions per genome per year,[30] nearly 60 times faster than *T pallidum*. Therefore, even *T pallidum* isolates with identical genomes do not necessarily indicate recent direct patient-to-patient transmission. Conversely, samples separated by even a very small number of SNPs are unlikely to share a recent common ancestor. Furthermore, because the potential transmission window of *T pallidum* might be as high as 2 years,[31,32] direct transmission cannot be excluded temporally for identical genomes collected within that period. This could ultimately limit our overall ability to deconvolute national or regional patterns of transmission.

See **Online** for appendix 2
See **Online** for appendix 3

Our study has several limitations, including the small number of samples compared to the total number of syphilis cases during the time period, and overrepresentation of samples from GBMSM in the national genome collection. The genomes represented 0·5% of the recorded number of syphilis cases in England during the study period and all samples referred to the National Reference Laboratory with sufficient treponemal DNA for sequencing. Although the available referral population might not be fully representative of syphilis in England due to regional variation in molecular testing and referral practices, all samples were collected and sequenced in the absence of any genetic relatedness information, so our genomic observations provide a snapshot of circulating English lineages. Future studies that focus on the systematic collection of samples from a higher proportion of cases, combined with improved sequence quality, will enable further insights into *T pallidum* transmission dynamics, and enable the fuller usefulness of sequence data to inform public health interventions.

Contributors

MAB, MJC, GH, MM, HF, and NRT conceptualised the study. MAB and LT contributed to methodology. MAB, LT, and HC did the formal analysis. MAB, LT, MJC, RP, and HC did the investigation. MJC, RP, ME, PF, MG, EEP, ES, JHV, KS, GH, MM, and HF provided resources. MAB, LT, HC, and KS did data curation. MAB wrote the original draft. MAB, LT, MJC, RP, HC, KS, GH, MM, HF, and NRT reviewed and edited the manuscript. MAB did the visualisation. MJC, KS, HF, and NRT supervised the trial. MAB, MJC, and MM did the project administration. KS, MM, HF, and NRT acquired funding. LT and HC accessed and verified the UKHSA data and MAB verified the underlying genomic data, the details of which have previously been reported. All authors had full access to all the data in the study (except for the detailed national patient metadata, which was restricted to UKHSA staff for patient confidentiality reasons) and accept responsibility to submit for publication. The corresponding authors had full access to all the data in the study (except for the detailed national patient metadata, which was restricted to UKHSA staff for patient confidentiality reasons), and had final responsibility for the decision to submit for publication.

Declaration of interests

We declare no competing interests.

Data sharing

Sequencing reads for all genomes used in this study have been previously published and described, and are available at the European Nucleotide Archive (https://www.ebi.ac.uk/ena) in BioProjects PRJEB28546, PRJEB33181, and PRJNA701499. All accessions, corresponding sample identifiers, and related metadata are available in appendix 2. Patient metadata for the UK genomes is available in pseudonymised form in appendix 3. UK shape files for Public Health region boundaries were downloaded from the UK Office for National Statistics, available at https://geoportal.statistics.gov.uk. The R code for all phylogenetic and statistical analysis and plotting is available in an Rnotebook along with all sample metadata (pseudonymised) and intermediate analysis files at https://doi.org/10.6084/m9.figshare.21543333.v1 and https://github.com/matbeale/Syphilis_Genomic_Epi_England_2022-23.

Acknowledgments

This research was funded in whole, or in part, by the Wellcome Trust (206194 and 220540/Z/20/A). MAB and NRT were supported by Wellcome funding to the Sanger Institute (206194 and 220540/Z/20/A). MM was funded by UK Research and Innovation and National Institute for Health and Care Research (COV0335; MR/V027956/1, NIHR200125) and the European and Developing Countries Clinical Trial Partnership (RIA2018D-249). Staff time for LT, MJC, RP, HC, KS, and HF and patient metadata retrieval and analyses were funded internally by UKHSA. We thank the sequencing team at the Wellcome Sanger Institute, and C Puethe and the Pathogen Informatics team for computational support, and additional technical staff involved in sample diagnostics, DNA extraction, and sample retrieval in laboratories at Public Health England (now UKHSA) and NHS laboratories (Birmingham, Brighton, Manchester, Leeds, and London), UK.

References

1 Fenton KA, Breban R, Vardavas R, et al. Infectious syphilis in high-income settings in the 21st century. *Lancet Infect Dis* 2008; 8: 244–53.
2 Centers for Disease Control and Prevention. 2017 sexually transmitted disease surveillance. Oct 15, 2018. https://www.cdc.gov/std/stats17/default.htm (accessed July 7, 2019).
3 Public Health England. Public Health England, National STI surveillance data tables 2020—table 1. 2021. https://www.gov.uk/government/statistics/sexually-transmitted-infections-stis-annual-data-tables (accessed May 12, 2022).
4 Centers for Disease Control and Prevention. Surveillance atlas of infectious diseases. 2017. https://www.ecdc.europa.eu/en/surveillance-atlas-infectious-diseases (accessed March 2, 2021).

5 Zhou Y, Li D, Lu D, Ruan Y, Qi X, Gao G. Prevalence of HIV and syphilis infection among men who have sex with men in China: a meta-analysis. *BioMed Res Int* 2014; **2014**: 620431.

6 Korenromp EL, Rowley J, Alonso M, et al. Global burden of maternal and congenital syphilis and associated adverse birth outcomes-estimates for 2016 and progress since 2012. *PLoS One* 2019; **14:** e0211720.

7 Tao Y, Chen MY, Tucker JD, et al. A nationwide spatiotemporal analysis of syphilis over 21 years and implications for prevention and control in China. *Clin Infect Dis* 2020; **70:** 136–39.

8 Public Health England. ISOSS congenital syphilis case review report: 2015 to 2020. 2021 https://www.gov.uk/government/publications/integrated-screening-outcomes-surveillance-service-isoss-annual-report/isoss-congenital-syphilis-case-review-report-2015-to-2020 (accessed May 12, 2022).

9 Furegato M, Fifer H, Mohammed H, et al. Factors associated with four atypical cases of congenital syphilis in England, 2016 to 2017: an ecological analysis. *Euro Surveill* 2017; **22:** 17-00750.

10 Pillay A, Liu H, Chen CY, et al. Molecular subtyping of *Treponema pallidum* subspecies pallidum. *Sex Transm Dis* 1998; **25:** 408–14.

11 Marra C, Sahi S, Tantalo L, et al. Enhanced molecular typing of treponema pallidum: geographical distribution of strain types and association with neurosyphilis. *J Infect Dis* 2010; **202:** 1380–88.

12 Grillova L, Jolley K, Šmajs D, Picardeau M. A public database for the new MLST scheme for *Treponema pallidum* subsp. *pallidum*: surveillance and epidemiology of the causative agent of syphilis. *PeerJ* 2019; **6:** e6182.

13 Grillová L, Bawa T, Mikalová L, et al. Molecular characterization of *Treponema pallidum* subsp. pallidum in Switzerland and France with a new multilocus sequence typing scheme. *PLoS One* 2018; **13:** e0200773.

14 Noda AA, Méndez M, Rodríguez I, Šmajs D. Genetic recombination in Treponema pallidum: implications for diagnosis, epidemiology, and vaccine development. *Sex Transm Dis* 2022; **49:** e7–10.

15 Sahi SK, Zahlan JM, Tantalo LC, Marra CM. A comparison of Treponema pallidum subspecies pallidum molecular typing systems: multilocus sequence typing vs. Enhanced centers for disease control and preventioning typing. *Sex Transm Dis* 2021; **48:** 670–74.

16 Liu D, He S-M, Zhu X-Z, et al. Molecular characterization based on MLST and ECDC typing schemes and antibiotic resistance analyses of *Treponema pallidum* subsp. *pallidum* in Xiamen, China. *Front Cell Infect Microbiol* 2021; **10:** 618747.

17 Beale MA, Marks M, Cole MJ, et al. Global phylogeny of Treponema pallidum lineages reveals recent expansion and spread of contemporary syphilis. *Nat Microbiol* 2021; **6:** 1549–60.

18 Beale MA, Marks M, Sahi SK, et al. Genomic epidemiology of syphilis reveals independent emergence of macrolide resistance across multiple circulating lineages. *Nat Commun* 2019; **10:** 3255.

19 Arora N, Schuenemann VJ, Jäger G, et al. Origin of modern syphilis and emergence of a pandemic *Treponema pallidum* cluster. *Nat Microbiol* 2016; **2:** 16245.

20 Lieberman NAP, Lin MJ, Xie H, et al. Treponema pallidum genome sequencing from six continents reveals variability in vaccine candidate genes and dominance of Nichols clade strains in Madagascar. *PLoS Negl Trop Dis* 2021; **15:** e0010063.

21 Taouk ML, Taiaroa G, Pasricha S, et al. Characterisation of Treponema pallidum lineages within the contemporary syphilis outbreak in Australia: a genomic epidemiological analysis. *Lancet Microbe* 2022; **3:** e417–26.

22 Nishiki S, Lee K, Kanai M, Nakayama SI, Ohnishi M. Phylogenetic and genetic characterization of Treponema pallidum strains from syphilis patients in Japan by whole-genome sequence analysis from global perspectives. *Sci Rep* 2021; **11:** 3154.

23 Dallman T, Ashton P, Schafer U, et al. SnapperDB: a database solution for routine sequencing analysis of bacterial isolates. *Bioinformatics* 2018; **34:** 3028–29.

24 Public Health England. Tracking the syphilis epidemic in England: 2010 to 2019. 2021. https://www.gov.uk/government/publications/tracking-the-syphilis-epidemic-in-england (accessed May 13, 2022).

25 Pinto M, Borges V, Antelo M, et al. Genome-scale analysis of the non-cultivable *Treponema pallidum* reveals extensive within-patient genetic variation. *Nat Microbiol* 2016; **2:** 16190.

26 Acheson P, McGivern M, Frank P, et al. An ongoing outbreak of heterosexually-acquired syphilis across Teesside, UK. *Int J STD AIDS* 2011; **22:** 514–16.

27 Glick SN, Morris M, Foxman B, et al. A comparison of sexual behavior patterns among men who have sex with men and heterosexual men and women. *J Acquir Immune Defic Syndr* 2012; **60:** 83–90.

28 Town K, Field N, Harris SR, et al. Phylogenomic analysis of *Neisseria gonorrhoeae* transmission to assess sexual mixing and HIV transmission risk in England: a cross-sectional, observational, whole-genome sequencing study. *Lancet Infect Dis* 2020; **20:** 478–86.

29 Williamson DA, Chow EPF, Gorrie CL, et al. Bridging of *Neisseria gonorrhoeae* lineages across sexual networks in the HIV pre-exposure prophylaxis era. *Nat Commun* 2019; **10:** 3988.

30 Sánchez-Busó L, Golparian D, Corander J, et al. The impact of antimicrobials on gonococcal evolution. *Nat Microbiol* 2019; **4:** 1941–50.

31 Schober PC, Gabriel G, White P, Felton WF, Thin RN. How infectious is syphilis? *Br J Vener Dis* 1983; **59:** 217–19.

32 Clark EG, Danbolt N. The Oslo study of the natural history of untreated syphilis; an epidemiologic investigation based on a restudy of the Boeck-Bruusgaard material; a review and appraisal. *J Chronic Dis* 1955; **2:** 311–44.

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/8605156

Genomic analysis of the host response to hepatitis B virus infection

Article *in* Proceedings of the National Academy of Sciences · May 2004

DOI: 10.1073/pnas.0401771101 · Source: PubMed

CITATIONS
700

READS
132

4 authors, including:

Robert Harry Purcell
National Institutes of Health
767 PUBLICATIONS **77,476** CITATIONS

Frank Chisari
The Scripps Research Institute
377 PUBLICATIONS **63,190** CITATIONS

SEE PROFILE

SEE PROFILE

All content following this page was uploaded by Robert Harry Purcell on 31 January 2016.

The user has requested enhancement of the downloaded file.

Genomic analysis of the host response to hepatitis B virus infection

Stefan Wieland*, Robert Thimme*[†], Robert H. Purcell[‡], and Francis V. Chisari*[§]

*Department of Molecular and Experimental Medicine, The Scripps Research Institute, La Jolla, CA 92037; and [‡]Hepatitis Viruses Section, Laboratory of Infectious Diseases, National Institutes of Health, Bethesda, MD 20892-8009

Contributed by Francis V. Chisari, March 12, 2004

Previous studies in hepatitis B virus (HBV)-infected humans and chimpanzees suggest that control of HBV infection involves the cells, effector functions, and molecular mediators of the immune response. The objective of the current study was to identify, in the liver of acutely HBV-infected chimpanzees, the spectrum of virus-induced and immune response-related genes that regulate the infection. The results demonstrate that HBV does not induce any genes during entry and expansion, suggesting it is a stealth virus early in the infection. In contrast, a large number of T cell-derived IFN-γ-regulated genes are induced in the liver during viral clearance, reflecting the impact of an adaptive T cell response that inhibits viral replication and kills infected cells, thereby terminating the infection.

The hepatitis B virus (HBV) is a noncytopathic hepatotropic DNA virus that causes acute chronic hepatitis and hepatocellular carcinoma (1). Viral clearance and disease pathogenesis during HBV infection are tightly associated with the appearance of a vigorous T cell response to all viral proteins (2, 3). In contrast, viral persistence and chronic hepatitis are associated with a markedly diminished HBV-specific T cell response (4, 5). CD8$^+$ T cells are the main immune effector cells during HBV infection, because viral clearance and liver disease are blocked by depletion of CD8$^+$ T cells in acutely infected chimpanzees (6). Furthermore, the onset of viral clearance in these animals is tightly associated with the appearance of virus-specific T cells (6) as well as CD3, CD8, and IFN-γ mRNA (6, 7) in the liver, indicating again that the adaptive T cell response plays a key role in this process.

Although cytolytic T cell functions certainly contribute to viral clearance, noncytolytic T cell functions also play a role, because adoptively transferred HBV-specific CD8$^+$ T cells inhibit viral replication in HBV transgenic mice by a noncytopathic IFN-γ-mediated mechanism (8), and because HBV DNA largely disappears from the liver and blood long before the peak of liver disease in acutely HBV-infected chimpanzees (6, 7). Furthermore, IFN-α/β-mediated mechanisms also inhibit HBV replication noncytopathically in transgenic mice (9), and they do so by inhibiting viral capsid assembly (ref. 10 and S.W. and F.V.C., unpublished observations). Interestingly, genomic analysis of the livers and hepatocyte cell lines from those mice demonstrated a close association between the antiviral effects of both IFN-γ and IFN-α/β and the induction of hepatocellular genes that might mediate the antiviral effects (11). These genes include GTP-binding proteins (e.g., GBP-1 and TGTP) known to inhibit other viral infections (12, 13), as well as components of the immunoproteasome (LMP2, LMP7, and MECL-1), IFN-stimulated protein 15 (ISG15), ubiquitin-specific protease 18 (Usp18), the chemokine IP-10, and the signal transducer and activator of transcription (STAT)-1 (11).

The current study was prompted by the desire to validate these findings in the setting of an acute HBV infection where the genomic changes occurring during viral entry, spread, and clearance can be uniquely identified. We did so by serially profiling the liver transcriptome in three acutely HBV-infected chimpanzees, searching especially for two distinct groups of cellular genes: those whose expression might correlate with the entry and expansion of the virus that might reflect the innate immune response, and those whose expression correlates with viral clearance reflecting the adaptive immune response that terminates infection.

Materials and Methods

Chimpanzees. Three healthy young adult HBV-seronegative chimpanzees (Ch1615, -1627, and -5835) were used in this study. The animals were handled according to humane use and care guidelines specified by the Animal Research Committees at the National Institutes of Health and The Scripps Research Institute. They were housed at Bioqual Laboratories (Rockville, MD), an American Association for Accreditation of Laboratory Animal Care International-accredited institution under contract to the National Institute of Allergy and Infectious Diseases. The animals were inoculated with 10^8 genome equivalents of a monoclonal HBV isolate (genotype *ayw*) contained in pooled serum from HBV transgenic mice (14). Before inoculation and weekly thereafter, blood was obtained by venipuncture and analyzed for serum alanine aminotransferase activity (sALT) as described (8). Six weeks after inoculation, Ch1615 and -1627 received three daily i.v. injections of either a humanized chimeric monoclonal anti-human CD4$^+$ antibody (cM-T412) or an irrelevant control antibody, respectively, as described (6). The course of infection, inflammatory infiltrate, and kinetics of viral clearance were not affected in these animals (6). Liver biopsy and DNA and RNA preparation protocols are available in *Supporting Materials and Methods*, which is published as supporting information on the PNAS web site.

Gene Expression Analysis. At selected time points, 100 ng of total liver RNA isolated from frozen liver biopsies was used to prepare cRNA by the reduced-volume RNA amplification protocol described by Baugh *et al.* (15). The cRNA was hybridized to high-density oligonucleotide arrays (HG-U133A Human GeneChips, Affymetrix, Santa Clara, CA), which interrogate the expression of ≈22,000 human genes. Primary image analysis was performed with GENECHIP VERSION 5.0 (Affymetrix). Relative gene expression levels were normalized throughout the entire data set by using algorithms in the DCHIP software package (16, 17). Genes not called "present" by GENECHIP at least at one time point were excluded from further analysis, and the remaining values were logarithm-transformed to base 10. Genes were identified whose expression correlated (by Pearson's correlation) either with the pattern of log-transformed intrahepatic HBV DNA or with the expression profile of specific prototype cellular genes described in the text. A Pearson correlation coefficient of ≥0.7 over all time points in all three chimpanzees was required for a gene to be further analyzed. Finally, selected genes had to be called "present" at the peak of intrahepatic HBV DNA levels or at the expression maxima of specific prototype genes, and they had to be called "induced" by GENECHIP at the same time point

---

Abbreviations: HBV, hepatitis B virus; STAT, signal transducer and activator of transcription; Ch*n*, chimpanzee *n*; HCV, hepatitis C virus; 2′5′OAS, 2′5′ oligoadenylate synthetase; sALT, serum alanine aminotransferase activity; NK, natural killer; TCR, T cell antigen receptor.

[†]Present address: Abteilung Innere Medizin II, Medizinische Universitätsklinik Freiburg, 79106 Freiburg, Germany.

[§]To whom correspondence should be addressed. E-mail: fchisari@scripps.edu.

© 2004 by The National Academy of Sciences of the USA

relative to the preinfection sample in all three animals. Genes were selected as "induced during viral clearance" if they were called "present" and "induced" relative to preinfection levels in weeks 12, 14, and 16 in Ch1627, 14 and 16 in Ch1615, and 14 and 18 in Ch5835. In addition, the expression level of these genes had to be higher during viral clearance than during the weeks before peak viremia in the same animal (i.e., weeks 4 and 6 in Ch1627 and 2 and 4 in Ch1615 and -5835). For Ch1627, HBV clearance-associated genes were grouped according to their induction kinetics into early, middle, and late genes based on the week (i.e., 12, 14, or 16) when peak induction was observed. Gene expression profiles were visualized by using GENESPRING VERSION 5.0 (Silicon Genetics, Redwood City, CA) by using gene expression values normalized to the 10th and 90th percentiles for each gene.

Results

Course of HBV Infection in Acutely Infected Chimpanzees. Ch1615, -1627, and -5835 were inoculated i.v. with 0.5 ml of HBV-transgenic mouse serum containing 1×10^8 genome equivalents of HBV DNA (6). As shown in Fig. 1, all three animals, including the CD4-depleted animal, became infected, developed acute hepatitis, and cleared the infection with remarkably similar kinetics (liver HBV DNA levels are shown in Table 2, which is published as supporting information on the PNAS web site). Viral spread in Ch1627 (Fig. 1*A*) and -5835 (Fig. 1*C*) was not accompanied by induction of CD3, IFN-γ, or 2′5′ oligoadenylate synthetase (2′5′OAS) mRNA, which was monitored by RNase protection analysis (Fig. 1), suggesting that the virus was not recognized by the innate or adaptive immune response in those animals at that time. In contrast, CD3, IFN-γ, and 2′5′OAS mRNA were all transiently induced in Ch1615 during week 5, coinciding with a transient spike of elevated sALT activity (Fig. 1*B*). All of these events were abruptly terminated by antibody-mediated CD4 depletion during week 6 (Fig. 1*B*, vertical arrow), suggesting that they reflected the atypically early onset of an adaptive immune response in this animal. This notion is further supported by the observation that, several weeks later, viral clearance in all animals coincided with the appearance of CD3, IFN-γ, and 2′5′OAS mRNA (Fig. 1), elevated sALT (Fig. 1), and HBV-specific T cells in the liver (ref. 6 and data not shown). Collectively, these results suggest that these T cells and their products contributed to viral clearance in these animals. Importantly, in all three animals, the viral DNA decreased 10- to 50-fold in the liver before the number of hepatitis B core antigen-positive hepatocytes began to decrease (Fig. 1), indicating that the decrease in viral DNA did not reflect the destruction of infected cells.

Liver Gene Expression Profiles in Acutely Infected Chimpanzees. The stars in Fig. 1 indicate the time points when liver gene expression profiling was performed in the animals. In this study, we attempted to identify two classes of genes: one group whose expression correlated with hepatic viral DNA content and therefore could reflect the initial (innate) host response to HBV in the liver, and a second group whose expression correlated with viral clearance and who are likely to be induced by the adaptive immune response to the virus.

Liver Genes Induced by HBV. The first group of genes displayed expression patterns that correlated (directly or inversely) with the amount of HBV DNA in the liver over the entire time course profiled. To reduce the chance of misinterpreting random fluctuations in gene expression in individual animals, we calculated the Pearson correlation between the expression of each gene and the amount of HBV DNA at each time point in all three animals, and we restricted our focus to genes whose changing expression levels correlated with the changing viral DNA content in the liver with a correlation coefficient of at least 0.7. Importantly, as shown in Fig. 2*A*, no genes fulfilled these criteria. Furthermore, no transcripts were uniformly induced or repressed if we limited our focus solely

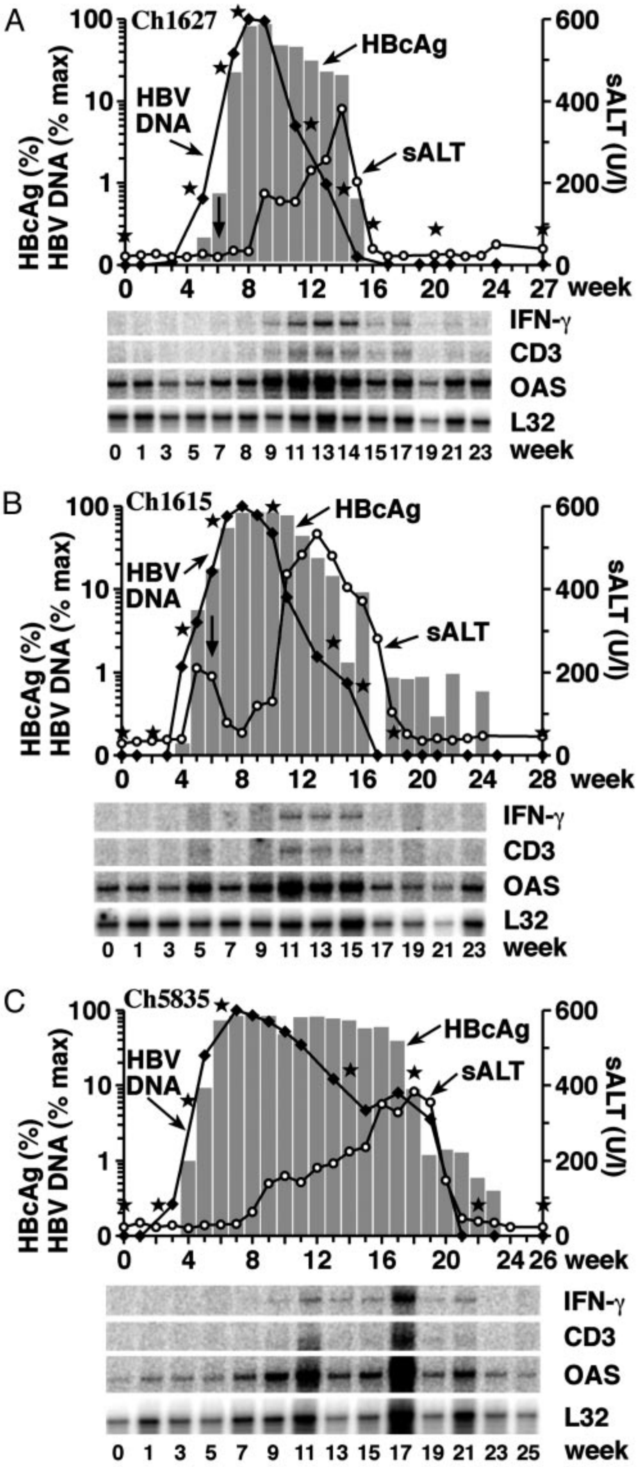


Fig. 1. Course of acute HBV infection in chimpanzees after experimental i.v. inoculation of 10^8 genome equivalents of HBV in week 0. (*A*) Ch1627 was also injected with an irrelevant control antibody in week 6. (*B*) Ch1615 was injected with a CD4-specific monoclonal antibody as described in the text. (*C*) Ch5835 did not receive any additional treatment. Intrahepatic HBV DNA (black squares) is expressed as a percentage (% max) of the peak HBV DNA levels in the liver of each animal (actual liver HBV DNA levels are presented in Table 2). Hepatitis B core antigen-positive (HbcAg) hepatocytes (gray bars) are expressed as percentage of the total number of hepatocytes. sALT (open circles) is expressed in units per liter. Intrahepatic expression profiles for IFN-γ, CD3, and 2′5′OAS were determined by RNase protection assay. Vertical arrows indicate the time of antibody injection. Stars indicate the time points for which gene expression profiling was performed.

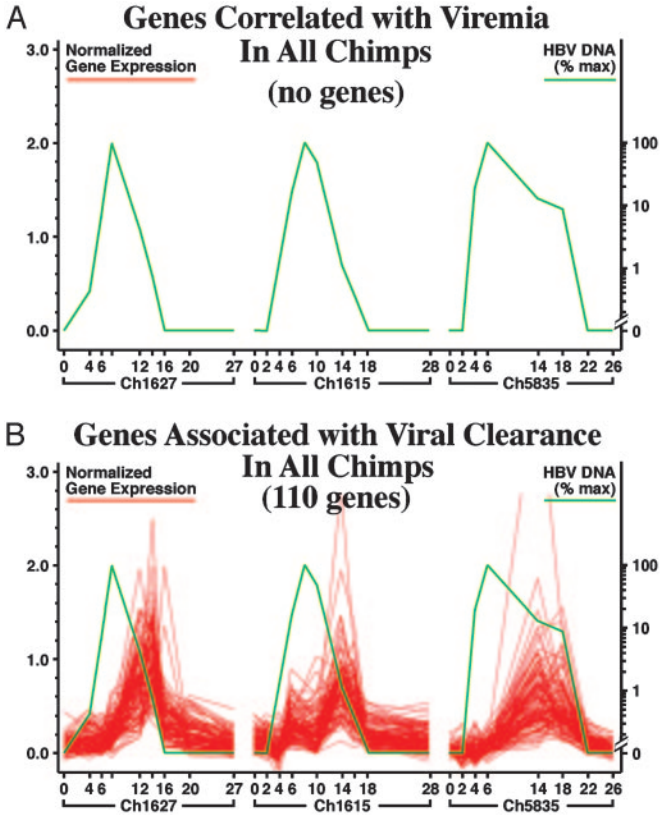


Fig. 2. Liver gene expression profile during acute HBV infection. (*A*) Genes correlated with viremia in all chimpanzees. No genes correlated positively or negatively with the level of intrahepatic HBV DNA during the time course of acute HBV infection in all three animals. (*B*) Liver gene expression profile associated with viral clearance in all three acutely HBV-infected chimpanzees. The gene identities are shown in Table 1 and the primary data in Tables 3 and 4. Green lines show the intrahepatic HBV DNA as a percentage (% max) of the corresponding peak levels. For optimal visualization, gene expression levels (red lines) are shown after normalization to the 10th and 90th percentiles for each gene. Values on the *x* axis represent weeks after inoculation with HBV.

to the lag phase of infection (weeks 0–2 in Ch1615 and -5835) or to the phase of logarithmic expansion (weeks 4–6 in Ch1615 and -5835 and weeks 4–7.5 in Ch1627) of the virus (data not shown). Because virtually 100% of the hepatocytes were infected in all three animals (see hepatitis B core antigen-positive hepatocytes in Fig. 1), the failure of the virus to induce cellular gene expression as it spread throughout the liver suggests that HBV does not induce an innate immune response at the site of infection.

Liver Genes Associated with Viral Clearance. To identify genes associated with viral clearance, we searched for transcripts that were either induced or suppressed when viral DNA disappeared from the liver. We identified 59 unique genes ("Selected as induced during viral clearance," Table 1, and indicated with a Y in column 5 of Table 3, which is published as supporting information on the PNAS web site) that were induced during viral clearance in all three animals (average peak fold change, 10.7), and none that were suppressed. The expression profile of these genes reflects a type 1 T cell response, because it was heavily weighted toward genes expressed by α/β T cells (e.g., T cell receptor β), γ/δ T cells (e.g., T cell receptor γ), the cytolytic effector molecule granzyme A, and many genes known to be induced by IFN-γ, which itself correlated with viral clearance (Fig. 1). The IFN-γ-induced genes include *STAT1*, MHC class I, MHC class II, immunoproteasome subunits *LMP2*, *PA28α* and -*β*, ubiquitin D, *E2L6*, tapasin, the chemokines monokine induced by IFN-γ (*MIG*), *IP-10*, and *RANTES*, the GTPase *GBP2*, and several other IFN-γ-induced genes [e.g., *IFI27* (18) and *IFI30* (19), and tryptophanyl-tRNA synthetase (20)], whose relevance to the T cell response and viral clearance is not immediately apparent. IFN-γ itself, however, is not in this list because it was scored present in only a few samples (data not shown and Table 4, which is published as supporting information on the PNAS web site), suggesting that the absolute level of intrahepatic IFN-γ expression (Fig. 1) was too low to be consistently detected by the microarrays used. 2′5′OAS, which is induced during viral clearance (Fig. 1), also is not part of this list, because its induction during viral clearance in Ch1615 was very brief (Table 4) and therefore did not satisfy our selection criteria. In addition, several genes not known to be inducible by IFN-γ (e.g., apolipoprotein L3, butyrophilin A3, cathepsin C, cofilin 1, complement component C1q, galectin 3, lysozyme, *Plac8*, superoxide dismutase, and Ig heavy and light chain alleles) were also induced during viral clearance.

Next, to identify viral clearance-associated genes that might have been excluded by our stringent selection criteria, we searched for genes whose expression correlated with the profiles of prototype genes, specifically T cell antigen receptor (*TCR*)*β*, *TCRγ*, *CD3D*, *IFN-γ*, *MIG*, *RANTES*, *IP-10*, and sALT, as indicated in columns 6–13 of Table 3. This increased the number of genes induced during viral clearance to 110, all of which are shown in Fig. 2*B* and identified in Tables 1 and 3.

This list contains additional T cell (*CD38*, *CD53*, and granzyme A), natural killer (NK) cell (*CD53* and granzyme K), B cell (*CD48* and *CD83*), macrophage (*CD68*), chemokine (*CXCL11* and *CXCR4*), GTPase (*GBP1*), complement (*C3a*), IFN-stimulated (*ISG20* and *IFI16*), and immunoproteasome (*MECL-1*) genes that were not included in the first-pass analysis. Importantly, 84 of these 110 genes were closely correlated with the TCRβ and -γ mRNA expression profiles (Table 3), 88 correlated with the induction profiles of *MIG*, *IP-10*, or *RANTES*, which are known to be highly induced by IFN-γ (21, 22), and 79 correlated with sALT levels in all three chimpanzees (Table 3 and Fig. 1). It is noteworthy that all of these genes returned to baseline when the viral DNA was eliminated (Fig. 2*B*). Because that correlation was not required for a gene to be designated as "clearance-associated," it strongly supports the notion that these genes were functionally related to viral clearance.

Sequential Induction of Clearance-Associated Liver Genes in Ch1627. Examination of the clearance-associated gene expression profile of Ch1627 (Figs. 2*B* and 3) reveals three distinct gene clusters whose expression peaked in week 12, 14, or 16 and were designated early, middle, and late clearance genes, accordingly (Fig. 3). The spacing of the sampled time points in the other chimpanzees did not permit sequential gene clusters to be identified in those animals. As shown in Table 1 and Fig. 3 (in blue), the early gene cluster consisted of 35 genes that included the TCRβ locus, several IFN-γ-induced genes including MHC class I and II genes, the TAP-binding protein (tapasin), the immunoproteasome components PA28α and -β, the chemokine *RANTES*, and IFN-γ-inducible protein 30, suggesting that they are mainly T cell-related and IFN-γ-induced genes that mediate antigen processing and presentation and inflammatory cell recruitment.

The middle gene cluster [Fig. 3 (shown in red) and Table 1] consisted of 73 genes, including *CD3D*, the TCRγ locus, a T cell activation marker (*CD38*), and the cytolytic effector granzymes A and K, as well as a T-LAK cell-originated protein kinase, a macrophage antigen (*CD68*), and a B cell membrane protein (*CD48*). Several additional IFN-γ-induced genes were also detected, including *STAT1*, the chemokines *IP-9*, *IP-10*, and *MIG*, the immunoproteasome component *MECL-1*, several ubiquitin-related genes, the GTPase *GBP1*, and its closely related family member GTPase guanylate-binding protein 2. Thus, the second wave of

Table 1. Liver genes associated with HBV clearance in all three chimpanzees

Gene/EST	Selected as induced during viral clearance	Peak expression* (Ch1627)	Induced during clearance in	
			HCV chimps†	HBVtg HCs‡
Adenosine deaminase		E		
Aldo-keto reduc. B1		M		
Annexin A2		E		
Apolipoprotein L 3	+	E		
ICB-1		M		
β-2-microglobulin		M		+
Butyrophilin 3 A3	+	M		
C3AR1		M		
Cat eye syndrome 1		M		
Cathepsin B		M		
Cathepsin Cd	+	M		
CD3-δ		M		
CD38		M		
CD48	+	M		
CD5	+	M	+	
CD53		M		
CD68		M		
CD74	+	E		
CD83	+	E		
CLN2	+	M		
Cofilin 1	+	E		
C1q	+	E	+	
CCL5 (RANTES)	+	E	+	
CXCL10 (IP-10)	+	M		+
CXCL11 (IP-9)		M		
CXCL9 (MIG)	+	M	+	+
CXCR4		M		
EMR1		M		
EST		E		
EST		E		
EST	+	M		
EST		M		
EST		E		
Fc-γ receptor I A1		M		
Fibrinogen-like 2	+	M		
Galectin 3	+	M		+
GBP1		M		+
GBP2	+	M	+	+
GM2	+	M		
Granzyme A	+	M	+	
Granzyme K		M		
HLA-A	+	E	+	+
HLA-B	+	E		+
HLA-C	+	E	+	
HLA-DMA	+	M	+	
HLA-DMB	+	M		
HLA-DPA1	+	E	+	
HLA-DPB1	+	M	+	
HLA-DQB1	+	E	+	
HLA-DRA	+	E	+	
HLA-DRB3	+	E		
HLA-E	+	E	+	+
IFI16		M		+
IFI27	+	M		
IFI30	+	E		
Isg20	+	E		+
IgHγ3	+	L		
Ig λ joining 3	+	L	+	
Ig superfamily 6		M		
Ig superfamily prot.	+	E		
IgE I receptor		E		
IL10R-α		M	+	
IAN4L1		E		
Karyopherin		M		
LAP3	+	M		+
Leupaxin		M		
Lysosome-assoc.-5	+	E		
Lysozyme	+	M		
MCM6		E		
Palmitoyl thioesterase 1		M		
PI3-kinase		E		
PTTG1		M		
PLAC8		M		+
PRC1		E		
Proteoglycan 1	+	M		
PSMB10 (MECL-1)		M	+	+
PSMB9 (LMP-2)	+	M	+	+
PSME1 (PA28 α)	+	E	+	+
PSME2 (PA28 β)	+	E	+	+
RAB20		M		
RAB27A		M		
RAB31		M		
RAC2	+	M		
RalGDS-like		M		
Regulatory factor X5	+	M		
RARRES3	+	M		
RhoGDI-β		M		
Ribonuclease T2		M		
Ribonuclease reduc. M2		M		
S100A10		E		
Ser/Thr kinase 6		M		
SOD2	+	M		
Solute carrier 7A7	+	M		
SP110		M		
STAT1	+	M	+	+
CHST6		M		
Tapasin	+	E		+
TCR β	+	E	+	
TCR γ	+	M		
TCR γ locus	+	M		
Thymidylate synthetase		M		
T-LAK prot. kinase		M		
GNMB		M		
TRIM22	+	M	+	
TrpRS	+	M	+	+
TYROBP	+	E	+	
Ubiquitin D	+	M		
UBE2C		M		
UBE2L6	+	M	+	+
Uncoupling prot. 2	+	E		

Table includes genes selected as induced during viral clearance (+) and/or correlated with prototype marker genes as described in the text. chimps, chimpanzees; HCs, hepatocytes; HBVtg, HBV transgenic; reduc., reductase; prot., protein.

*Peak expression early (E), middle (M), or late (L) during viral clearance in Ch1627.

[†] Induced during clearance in hepatitis C virus (HCV)-infected chimpanzees (23).

[‡] Induced by IFNβ and/or IFNγ in murine hepatocytes during clearance of HBV (11).

genes expressed during viral clearance in this animal reflects the amplification and diversification of the intrahepatic inflammatory response and the induction of antiviral activity within the infected cells.

Only two of the clearance-associated genes (i.e., Ig heavy and light chains) reached peak induction late in infection in this animal [Fig. 3 (shown in green) and Table 1], reflecting the expansion of a B cell infiltrate when the viral DNA was disappearing from the liver.

Discussion

The objective of this study was to identify liver genes that are transcriptionally regulated during HBV infection and are associated with either the entry and expansion of the virus or its immune-mediated clearance. The first group could include proviral genes that the virus might induce to optimize replication and antiviral genes that reflect the activation and/or effector functions of the innate immune response. The second group should include genes brought into the liver by cells of the adaptive immune response, plus

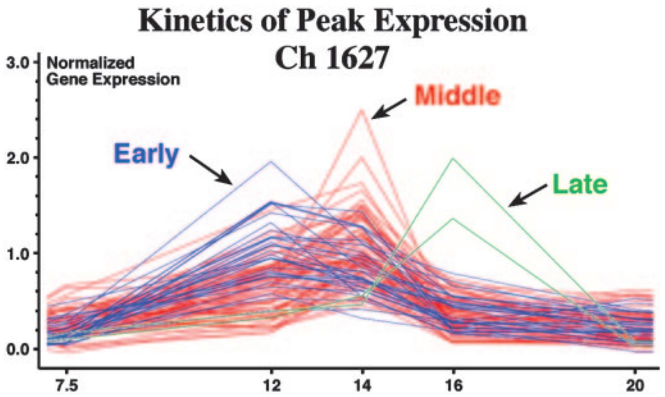


Fig. 3. Kinetics of liver gene expression during viral clearance in Ch1627. Early (blue), Middle (red), and Late (green) genes displayed maximal expression in weeks 12, 14, and 16, respectively.

hepatocellular genes that are activated by those cells and their products, e.g., IFN-γ.

These studies were modeled after a previous report from our laboratory (23) in which the genes associated with HCV spread and clearance were identified in the livers of chimpanzees that were acutely and chronically infected by HCV. Together, the results of the two studies display remarkable differences in the repertoire of genes that are induced during the infection by HBV in this study (no genes) and HCV in previous studies (23, 24) (27 genes, many of which are regulated by IFN-α/β), probably reflecting differences in the genome structure and the replication strategies of the two viruses and corresponding differences in their ability to induce an innate immune response. The results also reveal remarkable similarities in the repertoire of genes induced during viral clearance in both HBV and HCV infections (23, 24) and during T cell-mediated viral clearance in the HBV transgenic mouse model (11) (Table 1), probably reflecting similarities in the adaptive immune mechanisms responsible for their control. The results also illustrate that γ/δ T cells contribute to the inflammatory infiltrate during acute hepatitis B, and they identify a compact subset of candidate antiviral genes, some of which could mediate viral clearance by interrupting the HBV life cycle in infected cells.

IFN-α/β and -γ are known to inhibit viral replication in the liver of HBV transgenic mice (8–10, 25). Both cytokines are produced during the innate immune response, IFN-α/β primarily by infected cells and plasmacytoid dendritic cells (26, 27), and IFN-γ primarily by NK and NKT cells (28), which are abundant in the liver (28, 29). Thus, because we and others have shown that the expression of a large number of IFN-regulated genes is rapidly and consistently induced in the liver of HCV-infected chimpanzees and that the strength of their expression correlates with viremia (23, 24), we expected the same to occur in the HBV-infected animals. Surprisingly, not a single gene was induced or repressed by HBV in all three acutely infected animals. Furthermore, no genes were significantly up- or down-regulated in any of the animals in the lag phase of infection or the log phase of viral spread. These results imply that HBV did not induce an intrahepatic innate immune response that could be detected by gene chip analysis in all of the infected animals. The basis for this is not immediately clear. It may reflect the replication strategy of HBV whose DNA genome replicates within nucleocapsid particles (30) and thus is shielded from the cellular machinery that normally senses double-stranded RNA (31). It also implies that extracellular HBV virions and subviral products are not detected by the toll-like receptor system (32).

Many RNA viruses have developed elegant strategies to evade the antiviral effects of cytokines induced during the innate immune response (26). The need to develop these strategies exists for these viruses because they induce IFN-α/β in infected cells or IFN-γ by cells of the innate immune system. We suggest that HBV achieves the same end by not inducing these cytokines in the first place. An interesting consequence of this is that by not needing to evade IFN early in the infection, the virus remains susceptible to IFN later in the infection when the adaptive immune response begins (6, 33).

In contrast to the apparent invisibility of HBV to the innate sensing machinery, when CD3$^+$ MHC-restricted α/β receptor T cells appeared in the liver of the infected animals [Table 1 and Fig. 2*B* and 3 (shown in blue)], a temporal and functional cascade was initiated that correlated very strongly with the induction of IFN-γ and viral clearance. Because the IFN-γ-inducible chemokine MIG was induced immediately before the onset of clearance in all three animals (data not shown), it is likely that IFN-γ played a major role in inducing most of the viral clearance-associated genes detected in this study, including 2′5′OAS, which is known to be inducible by IFN-γ (34). By using Ch1627 as a prototype, the early appearance of TCRα/β receptor genes [Table 1 and Fig. 3 (shown in blue)] was accompanied by the induction of IFN-γ-regulated genes (e.g., MHC class I, MHC class II, immunoproteasome components *PA28*α and *PA28*β, and the TAP-binding protein tapasin), all of which may have facilitated antigen processing and presentation to the incoming T cells. In addition, the early induction of *RANTES*, *MIG*, and *IP-10* probably enhanced the recruitment of additional inflammatory cells into the liver. The remaining early genes underlie a wide array of cellular functions, some of which might contribute to viral clearance. For example, ISG20 is an IFN-inducible RNase that increases the resistance to infection by RNA viruses (35) and is also induced during IFN-mediated suppression of viral replication in HBV transgenic mouse hepatocyte cell lines (11). Thus, ISG20 may also be able to destroy HBV RNA in infected hepatocytes. Additional studies are needed to test this hypothesis.

The early peak was followed by a second group of genes [Table 1 and Fig. 3 (shown in red)] reflecting expansion of the inflammatory infiltrate to include γ/δ T cells. γ/δ T cells do not express CD4 or CD8, do not recognize peptide antigens, and are MHC-unrestricted (36), but they have been shown to produce IFN-γ and contribute to the clearance of other viral infections (36, 37). Thus, further investigation of γ/δ T cell immunobiology in HBV infection appears to be warranted.

The middle gene cluster [Table 1 and Fig. 3 (shown in red)] also displays further evidence of T and NK cells and macrophage activation in the liver, because *CD38* and the T cell and NK cell death effectors granzyme A and granzyme K peaked at this time, as did a large number of IFN-γ-regulated genes, consistent with the detection of IFN-γ mRNA and the decreasing viral DNA content of the liver (Fig. 1). For example, *STAT1*, which mediates IFN receptor signaling, was maximally induced at this time, as were several MHC class II genes, the chemokines *IP-10*, *MIG* and *IP-9*, the immunoproteasome components *LMP2* and *MECL-1*, ubiquitin D, and the ubiquitin-conjugating enzymes E2C and E2L6, all of which likely contribute to the inflammatory process. In addition, the IFN-γ-induced GTPases, GBP1 and -2, as well as *IFI27* and *IFI16*, peaked at this time. Interestingly, the murine homologues of *GBP1* and *IFI27* inhibit vesicular stomatitis virus (12) and Sindbis virus infection in mice (38), respectively, and they are induced in IFN-treated murine hepatocyte cell lines when HBV replication is suppressed (Table 1). Thus, further investigation of the role these genes play in viral clearance during HBV infection appears to be warranted.

Last, the middle gene cluster contains several other genes that could play a role in viral clearance and/or disease pathogenesis during HBV infection. Among these are members of the Ras-related family of small GTPases (i.e., *Rab20*, *-27a*, and *-31*) (reviewed in ref. 39) that might inhibit HBV replication by virtue of their ability to regulate vesicular transport (39–41). Cathepsin B is a cysteine protease that is released into the cytoplasm (42) during hepatocyte apoptosis where it might also target HBV. Karyopherin (importin-α) induction might also influence HBV replication because it is known to regulate HBV capsid nuclear import (43).

The last (late) wave consisted of Ig heavy and light chain genes [Table 1 and Fig. 3 (shown in green)], apparently reflecting an influx of B cells into the liver.

Although the results of this study elucidate and strongly underscore the role of the adaptive immune response and its products in viral clearance during acute HBV infection, they do not discriminate among genes that are expressed in the inflammatory cells or the hepatocytes. Although this should be clarified in future studies, some insight can be gained by comparing the current results with our previous analysis of the gene expression profiles of HBV transgenic murine hepatocyte cell lines in which HBV replication was suppressed by IFN-β and -γ *in vitro*. As shown in Table 1, a relatively compact panel of genes [*STAT1*, *IFI16*, *GBP1*, *GBP2*, *LMP2*, *MECL-1*, *PA28*α/β, *IP-10*, *MIG*, *ISG20*, leucine aminopeptidase 3, placenta-specific 8, tryptophanyl-tRNA synthetase (*TrpRS*), β-2-microglobulin, ubiquitin-conjugating enzyme *E2L6* (*UBE2L6*), tapasin, and MHC class I] was induced during viral clearance in all three chimpanzees and in IFN-β- and/or -γ-treated mouse hepatocytes (11), suggesting that they may function as effector molecules that suppress viral replication and contribute noncytolytically to the control of HBV infection. The generality and potential importance of this subset of genes are underscored by the fact that several of them are also induced in the liver during viral clearance in HCV-infected chimpanzees (23, 24).

Conclusion

The current results suggest that HBV acts like a stealth virus early in infection, remaining undetected and spreading until the onset of the adaptive immune response several weeks later. When MHC-restricted α/β T cells enter the liver and recognize antigen, they kill some of the infected cells and secrete IFN-γ, which induces the expression of a large number of genes that enhance antigen processing and presentation; recruit macrophages, NK cells, and γ/δ T cells that also produce IFN-γ; and amplify the process. Several of the IFN-γ-induced genes have antiviral activity in other systems, raising the possibility that they may also inhibit HBV replication. These highly coordinated cellular and molecular events continue until the infection is terminated, at which point they rapidly subside. Because many of the same genes are also associated with clearance of HCV, they represent a starting point for further elucidation of the cellular and molecular immunobiology of these infections.

We thank M. Shapiro and Dr. M. St. Claire (Bioqual, Rockville, MD) for animal care, C. Steiger for coordination and handling of biopsies, and R. Koch and R. Engle for technical assistance. We thank The Scripps Research Institute DNA core facility (Director Dr. S. Head) for RNA labeling, microarray hybridization, and data acquisition; Dr. A. Su (Genomics Institute of the Novartis Research Foundation, La Jolla, CA) for helpful advice; and Dr. R. Lanford (Southwest Foundation for Biomedical Research, San Antonio, TX) for critical reading of the manuscript. This study was supported by National Institutes of Health Grants AI20001 and CA76403 and contracts N01-AI-52705, N01-AI-45180, and N01-CO-56000, and The Sam and Rose Stein Charitable Trust. R.T. was supported by Grants TH 719/1-1 and TH 719/2-1 (Emmy Noether Program) from the Deutsche Forschungsgemeinschaft, Bonn, and by a postdoctoral training fellowship from the Cancer Research Institute, New York. This is manuscript number 16438-MEM from The Scripps Research Institute.

1. Chisari, F. V. & Ferrari, C. (1995) *Annu. Rev. Immunol.* **13,** 29–60.
2. Missale, G., Redeker, A., Person, J., Fowler, P., Guilhot, S., Schlicht, H. J., Ferrari, C. & Chisari, F. V. (1993) *J. Exp. Med.* **177,** 751–762.
3. Rehermann, B., Fowler, P., Sidney, J., Person, J., Redeker, A., Brown, M., Moss, B., Sette, A. & Chisari, F. V. (1995) *J. Exp. Med.* **181,** 1047–1058.
4. Rehermann, B., Chang, K. M., McHutchinson, J., Kokka, R., Houghton, M., Rice, C. M. & Chisari, F. V. (1996) *J. Virol.* **70,** 7092–7102.
5. Rehermann, B., Lau, D., Hoofnagle, J. H. & Chisari, F. V. (1996) *J. Clin. Invest.* **97,** 1655–1665.
6. Thimme, R., Wieland, S., Steiger, C., Ghrayeb, J., Reimann, K. A., Purcell, R. H. & Chisari, F. V. (2003) *J. Virol.* **77,** 68–76.
7. Guidotti, L. G., Rochford, R., Chung, J., Shapiro, M., Purcell, R. & Chisari, F. V. (1999) *Science* **284,** 825–829.
8. Guidotti, L. G., Ishikawa, T., Hobbs, M. V., Matzke, B., Schreiber, R. & Chisari, F. V. (1996) *Immunity* **4,** 25–36.
9. McClary, H., Koch, R., Chisari, F. V. & Guidotti, L. G. (2000) *J. Virol.* **74,** 2255–2264.
10. Wieland, S. F., Guidotti, L. G. & Chisari, F. V. (2000) *J. Virol.* **74,** 4165–4173.
11. Wieland, S. F., Vega, R. G., Muller, R., Evans, C. F., Hilbush, B., Guidotti, L. G., Sutcliffe, J. G., Schultz, P. G. & Chisari, F. V. (2003) *J. Virol.* **77,** 1227–1236.
12. Anderson, S. L., Carton, J. M., Lou, J., Xing, L. & Rubin, B. Y. (1999) *Virology* **256,** 8–14.
13. Carlow, D. A., Teh, S. J. & Teh, H. S. (1998) *J. Immunol.* **161,** 2348–2355.
14. Guidotti, L. G., Matzke, B., Schaller, H. & Chisari, F. V. (1995) *J. Virol.* **69,** 6158–6169.
15. Baugh, L. R., Hill, A. A., Brown, E. L. & Hunter, C. P. (2001) *Nucleic Acids Res.* **29,** E29.
16. Schadt, E. E., Li, C., Ellis, B. & Wong, W. H. (2001) *J. Cell. Biochem. Suppl.* **37,** 120–125.
17. Li, C. & Hung Wong, W. (2001) *Genome Biol.* **2,** RESEARCH0032.
18. Parker, N. & Porter, A. C. (2004) *BMC Genomics* **5,** 8.
19. Watts, C. (2001) *Science* **294,** 1294–1295.
20. Liu, J., Shue, E., Ewalt, K. L. & Schimmel, P. (2004) *Nucleic Acids Res.* **32,** 719–727.
21. Kakimi, K., Lane, T. E., Wieland, S., Asensio, V. C., Campbell, I. L., Chisari, F. V. & Guidotti, L. G. (2001) *J. Exp. Med.* **194,** 1755–1766.
22. Losana, G., Bovolenta, C., Rigamonti, L., Borghi, I., Altare, F., Jouanguy, E., Forni, G., Casanova, J. L., Sherry, B., Mengozzi, M., *et al.* (2002) *J. Leukocyte Biol.* **72,** 735–742.
23. Su, A. I., Pezacki, J. P., Wodicka, L., Brideau, A. D., Supekova, L., Thimme, R., Wieland, S., Bukh, J., Purcell, R. H., Schultz, P. G., *et al.* (2002) *Proc. Natl. Acad. Sci. USA* **99,** 15669–15674.
24. Bigger, C. B., Brasky, K. M. & Lanford, R. E. (2001) *J. Virol.* **75,** 7059–7066.
25. Guidotti, L. G., Borrow, P., Hobbs, M. V., Matzke, B., Gresser, I., Oldstone, M. B. & Chisari, F. V. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 4589–4594.
26. Katze, M. G., He, Y. & Gale, M., Jr. (2002) *Nat. Rev. Immunol.* **2,** 675–687.
27. Siegal, F. P., Kadowaki, N., Shodell, M., Fitzgerald-Bocarsly, P. A., Shah, K., Ho, S., Antonenko, S. & Liu, Y. J. (1999) *Science* **284,** 1835–1837.
28. Bendelac, A., Rivera, M. N., Park, S. H. & Roark, J. H. (1997) *Annu. Rev. Immunol.* **15,** 535–562.
29. Eberl, G., Lees, R., Smiley, S. T., Taniguchi, M., Grusby, M. J. & MacDonald, H. R. (1999) *J. Immunol.* **162,** 6410–6419.
30. Seeger, C. & Mason, W. S. (2000) *Microbiol. Mol. Biol. Rev.* **64,** 51–68.
31. Samuel, C. E. (2001) *Clin. Microbiol. Rev.* **14,** 778–809.
32. Hertzog, P. J., O'Neill, L. A. & Hamilton, J. A. (2003) *Trends Immunol.* **24,** 534–539.
33. Wieland, S. F., Spangenberg, H. C., Thimme, R., Purcell, R. H. & Chisari, F. V. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 2129–2134.
34. Guidotti, L. G., Morris, A., Mendez, H., Koch, R., Silverman, R. H., Williams, B. R. & Chisari, F. V. (2002) *J. Virol.* **76,** 2617–2621.
35. Espert, L., Degols, G., Gongora, C., Blondel, D., Williams, B. R., Silverman, R. H. & Mechti, N. (2003) *J. Biol. Chem.* **278,** 16151–16158.
36. Hayday, A. C. (2000) *Annu. Rev. Immunol.* **18,** 975–1026.
37. Wang, T., Scully, E., Yin, Z., Kim, J. H., Wang, S., Yan, J., Mamula, M., Anderson, J. F., Craft, J. & Fikrig, E. (2003) *J. Immunol.* **171,** 2524–2531.
38. Labrada, L., Liang, X. H., Zheng, W., Johnston, C. & Levine, B. (2002) *J. Virol.* **76,** 11688–11703.
39. Seabra, M. C., Mules, E. H. & Hume, A. N. (2002) *Trends Mol. Med.* **8,** 23–30.
40. Trambas, C. M. & Griffiths, G. M. (2003) *Nat. Immunol.* **4,** 399–403.
41. Pereira-Leal, J. B. & Seabra, M. C. (2001) *J. Mol. Biol.* **313,** 889–901.
42. Guicciardi, M. E., Deussing, J., Miyoshi, H., Bronk, S. F., Svingen, P. A., Peters, C., Kaufmann, S. H. & Gores, G. J. (2000) *J. Clin. Invest.* **106,** 1127–1137.
43. Rabe, B., Vlachou, A., Pante, N., Helenius, A. & Kann, M. (2003) *Proc. Natl. Acad. Sci. USA* **100,** 9849–9854.

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/233948663

Genomic correlates of variability in immune response to an oral cholera vaccine

Article *in* European Journal of Human Genetics · December 2012

DOI: 10.1038/ejhg.2012.278 · Source: PubMed

CITATIONS
9

READS
57

9 authors, including:

Partha Majumder
National Institute of Biomedical Genomics
325 PUBLICATIONS **15,397** CITATIONS
SEE PROFILE

Neeta Sarkar-Roy
National Institute of Biomedical Genomics
19 PUBLICATIONS **822** CITATIONS
SEE PROFILE

Herman F Staats
Duke University Medical Center
136 PUBLICATIONS **7,035** CITATIONS
SEE PROFILE

Thandavarayan Ramamurthy
National Institute of Cholera and Enteric Diseases
524 PUBLICATIONS **22,980** CITATIONS
SEE PROFILE

All content following this page was uploaded by Partha Majumder on 29 September 2016.

The user has requested enhancement of the downloaded file.

European Journal of Human Genetics (2013) **21,** 1000–1006
© 2013 Macmillan Publishers Limited All rights reserved 1018-4813/13
www.nature.com/ejhg

ARTICLE

Genomic correlates of variability in immune response to an oral cholera vaccine

Partha P Majumder[*,1,2], Neeta Sarkar-Roy[1,2], Herman Staats[3], T Ramamurthy[4], Sujit Maiti[2,5], Goutam Chowdhury[4], Carol C Whisnant[6], K Narayanasamy[7,8] and Diane K Wagener[6]

Cholera is endemic to many countries. Recent major outbreaks of cholera have prompted World Health Organization to recommend oral cholera vaccination as a public-health strategy. Variation in percentage of seroconversion upon cholera vaccination has been recorded across populations. Vaccine-induced responses are influenced by host genetic differences. We have investigated association between single-nucleotide polymorphic (SNP) loci in and around 296 immunologically relevant genes and total anti-lipopolysaccharide (LPS) antibody response to a killed whole-cell vaccine, comprising LPS from multiple strains of *Vibrio cholerae*. Titers derived from standard vibriocidal assays were also analyzed to gain further insights on validated SNP associations. Vaccination was administered to 1000 individuals drawn from India. Data on two independent random subsets, each comprising ~500 vaccinees, were used for discovery of genomic associations and validation, respectively. Significant associations of four SNPs and haplotypes in three genes (*MARCO, TNFAIP3* and *CXCL12*) with AR were discovered and validated, of which two in *TNFAIP3* and *CXCL12* were also significantly associated with immunity (fourfold increase in vibriocidal titers). *CXCL12* is a neutrophil and lymphocyte chemoattractant that is upregulated in response to *V. cholerae* infection. LPS in the vaccine possibly provides signals that mimic those of the live bacterium. *TNFAIP3* promotes intestinal epithelial barrier integrity and provides tight junction protein regulation; possible requirements for adequate response to the vaccine. LPS is a potent activator of innate immune responses and a ligand of *MARCO*. Variants in this gene have been found to be associated with LPS response, but not with high vibriocidal titer level.

European Journal of Human Genetics (2013) **21,** 1000–1006; doi:10.1038/ejhg.2012.278; published online 19 December 2012

Keywords: anti-LPS antibody assay; vibriocidal assay; SNP; haplotype; association

INTRODUCTION

Major outbreaks of cholera, leading to several thousand deaths, in Zimbabwe in 2009 and in Haiti in 2010, prompted the World Health Organization to emphasize the administration of oral cholera vaccination.[1–4] Cholera is caused by toxicogenic strains of the Gram-negative bacterium *Vibrio cholerae*. Moderate to severe watery diarrhea, that appears within 1–5 days of ingestion of contaminated food or water, rapidly leads to dehydration and, if left untreated, to death. It is a significant cause of mortality in developing countries,[5] and is still endemic in many countries, including India.[6] Unfortunately, 13% of the world's population still lacks access to safe water.[7]

V. cholerae, has two major serogroups – O1 (serotypes Inaba and Ogawa) and O139 – that are responsible for most cholera cases and differ in the lipopolysachharide (LPS) structures of the outer membrane.[8] Immunity to cholera can be acquired.[9] WHO advocates vaccination for cholera.[10] The efficacy of currently available vaccines varies across populations; protective efficacy is unlikely to last more than 3 years.[11] A single-dose cholera vaccine that induces long-term immunity for those at risk remains an unfulfilled goal.[12] Persons of blood group O are more susceptible to cholera than non-O blood group persons.[13–15] However, the association between efficacy of cholera vaccines and O blood group has been inconsistent–negative in Bangladesh,[13] positive in Indonesia[16] and no association in India.[17] Inter-individual variability in immune response to vaccines impedes the control of infectious diseases. Primary failure, as assessed by postvaccination antibody levels, of vaccines is a problem; 2–10% for measles vaccine,[18,19] 5–20% for hepatitis B vaccine[20,21] and 23–40% for typhoid vaccine.[22,23] Data are unavailable for cholera vaccines. Twin studies have shown high heritability for antibody response (AR) to measles (89%), mumps (39%), rubella (46%), hepatitis B (61–77%), oral polio (60%), tetanus (44–64%), diphtheria (49%), *Haemophilus influenzae* type B (Hib) (51%) and other vaccines.[24] Polymorphic variants in HLA and other genes–such as, *IL1B, IL4R, IL6, IL10* and *TNF*–are associated or linked with AR to several vaccines.[18,25–30] Although measles, mumps and rubella vaccines are all attenuated-live vaccines that are administered simultaneously, the responses induced by these vaccines are influenced differently by host genetics.[24] Vietnam is the only country in the world with large cholera-endemic tracts where an oral vaccine for cholera has been widely used for a long time (since 1997).[31] Considerable variation in response among vaccinees was noted in a clinical trial of this vaccine[31] conducted in a non-endemic area of Vietnam (Son La province), although >90% of vaccinees seroconverted. However, in respect of the same vaccine, only ~50% of vaccinees seroconverted in the endemic area of West Bengal, India.[17]

---

[1]National Institute of Biomedical Genomics, Kalyani, India; [2]TCG-ISI Centre for Population Genomics (CpG), Kolkata, India; [3]Duke University Medical Centre, Durham, NC, USA; [4]National Institute of Cholera and Enteric Diseases, Kolkata, India; [5]University of Western Ontario, Ontario, London, Canada; [6]RTI International, Research Triangle Park, Durham, NC, USA; [7]The Centre for Genomic Applications (TCGA), New Delhi, India; [8]SphaeraPharma, Manesar, India

*Correspondence: Professor PP Majumder, National Institute of Biomedical Genomics, Netaji Subhas Santorium (2nd Floor), Kalyani 741251, India. Tel: +91 33 25892151; Fax: +91 33 25892150; E-mail: ppm1@nibmg.ac.in

Received 10 July 2012; revised 23 October 2012; accepted 20 November 2012; published online 19 December 2012

Vaccine response is not wholly determined by the genetic make-up of vaccine recipients, but also by the nature of microbial antigens and adjuvants.[32] Knowledge of key immunogenetic associations can be used to design a vaccine that circumvents immunogenetic restrictions.[33] Insights into the impact of genetic variation in vaccinees on AR has been obtained for some vaccines.[24,34,35] In this study, we have investigated, for the first time, whether variation in the genomic backgrounds of vaccinees contributed to variation in immunological response to the killed whole-cell bivalent oral cholera vaccine, which has been introduced in India in 2010.

METHODS

Study participants

Individuals ($n=1000$), unrelated at least to the first-cousin level based on family history report, aged 14 years or older, inhabiting a socio-economically depressed locality of Kolkata (formerly, Calcutta), India, were recruited using a statistical sampling design,[34] with institutional ethical approval and written individual informed consent. Exclusions included: (a) pregnant and lactating women and (b) based on self-report, individuals who had ever been diagnosed with cholera or had suffered from diarrhea or vomiting during the week preceding recruitment. The sample of 948 vaccinees, whose data passed all the quality checks (explained later), was randomly split into two half-samples (HSs); one used for association-discovery and the other for association-validation.

Vaccination and collection of blood samples

With approval of the Drug Controller General of India, a two-dose vaccine (see online Supplementary Table S1 for composition) procured from Shantha Biotechnics, Hyderabad, India, orally administered 14 days apart, was used. The vaccine was stored and transported at 4–8 °C before administration. From each vaccinee, a blood sample was collected immediately before vaccination (Day 0 sample) and 28 days postvaccination (Day 28 sample).

Assessment of antibody level

Antibody level was assessed using two different methods as described below.

Total anti-LPS AR assay. Serum anti-*V. cholerae* (Inaba 569) LPS–that is, total anti-LPS antibody–was assayed using the Bio-Plex (Bio Rad, Hercules, CA, USA) platform (hereafter referred to as the 'Bioplex assay'). *V. cholerae* O1 (Inaba 569) LPS (List Biological, Cat. No.231) was conjugated to Bio Rad beads (COOH (042)) by Solulink (San Diego, CA, USA). Day 0 and Day 28 human sera were incubated with LPS beads with anti-*V. cholerae* LPS antibody and responses were detected using a cocktail of anti-human IgG-PE, IgA-PE and IgM-PE. A pooled serum with a high anti-LPS IgG/IgA/IgM response was used to create a standard curve that was assigned an arbitrary value of anti-LPS Ig ELISA units (EU)/ml. Responses (Day 0 and Day 28) in individual subjects were compared with the standard curve to calculate anti-LPS Ig EU/ml.

Vibriocidal assay. This assay was performed with *V. cholerae* O1 Inaba (OS-418), Ogawa (MAK757) and uncapsulated O139 (MO-10T4) strains using sera collected during pre- and postvaccine trials following published methods.[36] Commercially available guinea-pig serum (Rockland, Gilbertsville, PA, USA) was used as a source of complement. The sera (100 μl) were added to 100 μl of PBS in the first well to give twofold dilution and the subsequent dilutions were made reciprocally up to 4800. Reference rabbit antiserum against *V. cholerae* O1 Inaba, Ogawa and O139 were included in each set of assay as controls. In every batch of assay, sera obtained from a healthy volunteer and a high-titer antiserum obtained from one of the volunteers in this study were included as negative and positive controls, respectively.

Candidate genes and SNPs

We selected 296 autosomal genes from immunological pathways as candidates for this association study. Two thousand four hundred and eighty-eight SNPs from these genes (Supplementary Table S2 online), including about 2-kb upstream and and 1-kb downstream regions, were chosen from the HapMap database (www.hapmap.org) using a statistical protocol[37] that maximized 'informativeness' of selected SNPs. Individuals belonging to the first HS were genotyped at all the 2488 loci. Genotypes at those loci that showed statistically significant association with AR in the first HS were determined in individuals belonging to the second HS for validation of the discovered associations. Genotyping was done using the Illumina iSCAN (San Diego, CA, USA) platform using Golden Gate assays following protocols recommended by the manufacturer.

Statistical analysis

AR to vaccination was measured as the ratio (fold-change) of Day 28 postvaccination to prevaccination (Day 0) antibody levels (EU/ml). Individuals on whom AR could not be properly assayed (5%) were removed from analyses. Curation of genotype data included removal of (a) loci with $<90\%$ call rate (2% of loci), (b) non-polymorphic loci with MAF<0.05 (12.7% of loci with call rate $>90\%$), (c) loci significantly deviant from HWE (1% of polymorphic loci), and (d) individuals with $<90\%$ of loci with valid genotype calls (0.7%). After data-curation, 948 individuals and 2036 SNPs were included for further analysis. Analysis of variance (ANOVA) was performed to test equality of mean values of *ln*AR ($\log_e$-transformation was used to induce Normality) among subgroups of vaccinees. To avoid detection of spurious associations arising from systematic ancestry differences, population stratification analysis was performed using Eigenstrat.[38] Instead of using an arbitrary threshold to classify a vaccinee as a 'responder' or a 'non-responder', we sorted the vaccinees by ascending order of their AR and grouped the vaccinees into five pentile (20 percentile) groups (PGs). It is expected that the frequency of an allele that is significantly associated with response should increase (or decrease) from low responders to high responders. Detection of allelic association was done by estimating the correlation coefficient (r) between the MAF at each locus and mean AR of vaccinees belonging to the (PGs) and testing H_0: $r=0$ using a permutation test. The false discovery rate procedure[39] was used to identify statistically significant values of r at the 5% level of significance. As correction for multiple testing when loci are in significant linkage disequilibrium (LD) remains a major statistical problem,[40,41] for single-locus association analysis, we circumvented the problem by selecting from each gene only the highly informative and minimally associated SNPs.[37] Four hundred and sixty-nine SNPs were thus selected; data on the remaining loci were kept aside for haplotype association analysis. Each significant allelic association was revalidated by confirming that the null hypothesis that the mean *log*AR values among the genotype classes are equal was rejected, using ANOVA. The standard regression procedure was used to estimate effect size (slope, β) of each significantly associated SNP. SD of effect size was estimated by bootstrapping 100 times; that is, by randomly choosing the appropriate number of samples with replacement from the original data and estimating β from the bootstrap sample and then calculating the empirical SD of the 100 estimates of β. Haplotypes were inferred using Phase.[42] Haplotype association analysis was performed using a permutation algorithm similar to that for single SNP markers.

RESULTS

Demographic characteristics of vaccinees

No statistically significant differences in the demographic characteristics of the vaccinees (Table 1) were noted between the association discovery and validation HSs in respect of age (mean ages were 34.24 ± 0.60 years and 34.36 ± 0.58 years; $p=0.883$), proportions of vaccinees belonging to the two genders ($p=0.602$) or the two religious groups ($p=0.21$).

Inter-individual variation in pre- and post-vaccination antibody levels and AR

Large inter-individual variations of pre- and postvaccination antibody levels and also of AR to vaccination (fold-change) were noted (Table 2 and Supplementary Table S3 online). Mean values between the two HSs were not significantly different (Supplementary Table S3).

Table 1 Demographic characteristics of the vaccine recipients

		Half-sample 1 (Association discovery sample)		*Half-sample 2 (Association validation sample)*	
		Gender		*Gender*	
Religion	*Age group (years)*	*Male*	*Female*	*Male*	*Female*
Muslim					
	<30	74	21	47	30
	30–49	51	37	52	31
	≥50	11	11	11	14
	Total	136	69	110	75
Hindu					
	<30	45	73	53	56
	30–49	56	52	71	71
	≥50	21	22	15	23
	Total	122	147	139	150
Total		258	216	249	225

Table 2 Mean ± SD of antibody levels and response using bioplex and vibriocidal assays

Assay	*Variable*	n	*Mean*	*SD*
Bioplex[a]	Day 0 (EU/ml)	948	22.54	37.95
	Day 28 (EU/ml)	948	76.89	114.88
	Fold change	948	5.42	12.41
	ln(Fold change)	948	1.14	0.84
Vibriocidal titer	Day 0	935	58.32	111.97
	Day 28	936	127.66	179.78
	Fold change	935	14.74	54.55

[a] Total anti-LPS antibody level.

Table 3 Comparison of antibody response estimated using the vibriocidal and bioplex assays

	Fold change–bioplex assay					
Fold change–vibriocidal assay	*<2*	*2–3*	*3–4*	*4–8*	*≥8*	*Total*
<2	123	67	18	17	9	234
2–4	100	55	34	41	29	259
4–8	39	43	18	47	16	163
8–32	35	29	28	28	39	159
≥32	26	20	13	30	31	120
Total	323	214	111	163	124	935

χ^2 value = 135.8 (d.f. = 16, $p < 0.0001$).
Contingency coefficient = 0.356 ($p < 0.0001$).

The correlation between baseline and Day 28 antibody levels is positive and moderate (0.449), but significant ($p < 0.0001$).

Antibody response

AR-estimates obtained from the two separate assays were significantly ($p < 0.0001$) correlated (Table 3). As, by its very nature, the vibriocidal assay is less quantitative than the bioplex assay, in further analyses, unless specifically mentioned, AR as measured by the bioplex assay was used. Mean values of *ln*AR for the two HSs did not differ significantly ($p = 0.564$) and were, 1.128 ± 0.038 and 1.160 ± 0.039, respectively. None of the possible covariates–religion, age and gender–had a statistically significant effect on *ln*AR, either individually or interactively (*p*-values for main effects and interactions were all >0.05). Therefore, all data were pooled in subsequent analyses.

Population stratification

No pair of vaccinees in the first HS shared >90% of the 2036 loci with identical genotypes; i.e., there was no evidence of cryptic relatedness among the vaccinees. There was also no evidence of population stratification as all individuals formed a single cluster in the scatterplot of the first three principal components (Figure 1).

Antibody levels and response in vaccinees belonging to the five PGs

Mean values of antibody levels and response were nearly equal for each PG between the HSs (Supplementary Table S4 online), but differed significantly ($P < 0.05$) among the PGs (Table 4). (Mean values of fold-change based on the vibriocidal assay were consistently higher than those based on the bioplex assay.) Mean AR increased almost linearly from PG 1 through 4, after which there was a steep increase (Figure 2), which was also observed for values based on vibriocidal assay. The steep increase of mean AR in PG-5 is primarily because of the presence of some individuals who exhibited very high AR.

Association results: discovery and validation

In the discovery HS, correlations between mean AR and MAFs in PGs were found to be statistically significant at 159 SNP loci on 93 genes (of the 296 genes considered in this study; Supplementary Table S5 online). The associations of these 159 SNPs could be validated in HS-2 only for seven SNPs that are located in and around five genes (rs17180481 and rs17180600 in the flanking region and intron, respectively, of *MARCO* (2q14.2); rs598493 in an intron of *TNFAIP3* (6q23.3); rs3012694 and rs266087 in introns of *RFX3* (9p24.2); rs266087 in an intron of *CXCL12* (10q11.21); and rs6567272 in an intron of *TNFRSF11A* (18q22.1)). Mean AR for vaccinees belonging to the three genotypes for each of these seven SNP loci differed significantly for only the four SNPs in and around *MARCO, TNFAIP3* and *CXCL12* (rs17180481: $p = 0.023$, effect size = 1.8 ± 0.3; rs17180600: $p = 0.026$, effect size = 1.6 ± 0.3; rs598493: $p = 0.001$, effect size = 3.9 ± 0.8; rs266087: $p = 0.014$, effect size = 2.4 ± 0.6), but not for the remaining three. The MAF values in PGs are provided in Figure 3 for these four SNPs.

Haplotype association

For these three genes, data on additional SNPs, which were in significant LD with the SNPs included in the association analysis, were used to reconstruct haplotypes and estimate haplotype frequencies among the vaccinees. Figure 4 provides, for *MARCO, TNFAIP3* and *CXCL12,* the frequency of the haplotype that showed the maximum association with AR by PG. The strength of association (i.e., slope of the line) with AR increased significantly for haplotypes compared with the single SNPs in these genes, indicating that these three genes are important modulators of AR to the vaccine.

Concordance between anti-LPS antibody and vibriocidal responses

For the four loci that were validated to be significantly associated with AR, the proportions of individuals who elicited fourfold or higher

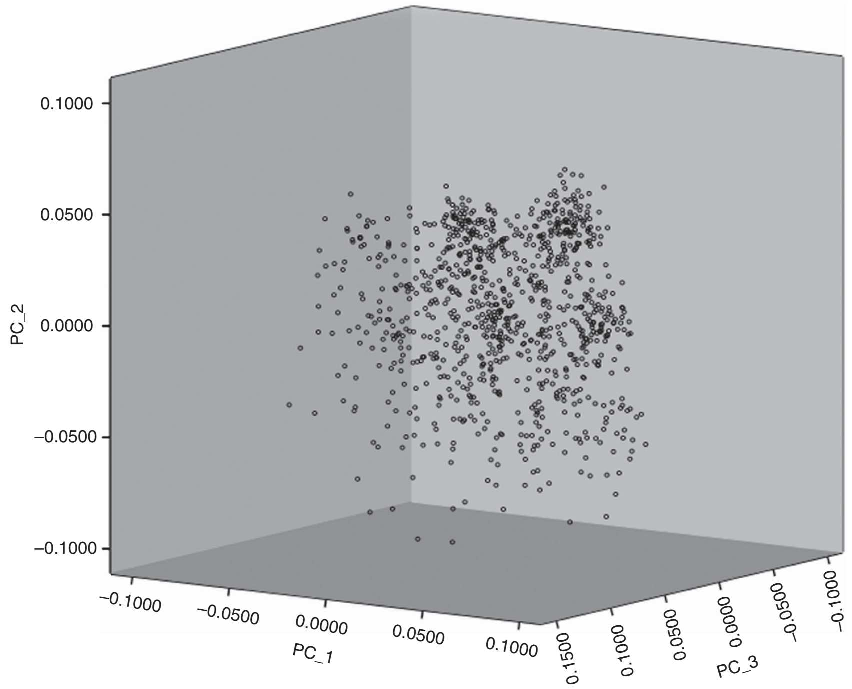


Figure 1 Scatterplot of vaccinees based on first three principal components extracted using data on 2036 SNPs show a single cluster and no evidence of population stratification.

Table 4 Means and SDs of total anti-LPS antibody levels on days 0 and 28, and antibody response, for the five pentile groups based on antibody response

Variable	*Pentile group*	*n*	*Mean*	*SD*
Baseline (Day 0) Antibody level	1	190	41.58	71.51
	2	190	24.10	26.66
	3	190	17.92	21.03
	4	190	17.16	17.56
	5	188	11.82	10.46
	Total	948	22.54	37.95
Day 28 antibody level	1	190	52.61	86.82
	2	190	44.99	52.88
	3	190	46.39	54.48
	4	190	72.73	77.74
	5	188	168.70	190.30
	Total	948	76.89	114.88
Antibody response	1	190	1.31	0.15
	2	190	1.84	0.17
	3	190	2.60	0.28
	4	190	4.20	0.73
	5	188	17.29	24.45
	Total	948	5.42	12.41

titer increase in the vibriocidal assay (that is likely to be suggestive of protective immunity,[43,44] although true protective immunity must be ascertained by vaccine challenge studies) significantly differed among the three genotypes at the locus only for two (rs598493 in *TNFAIP3*; $p=0.001$ and rs266087 in *CXCL12*; $p=0.014$) of these four loci (Figure 5). Thus, although *MARCO* was found to be significantly associated with AR, it was not found to be associated with high vibriocidal titer levels.

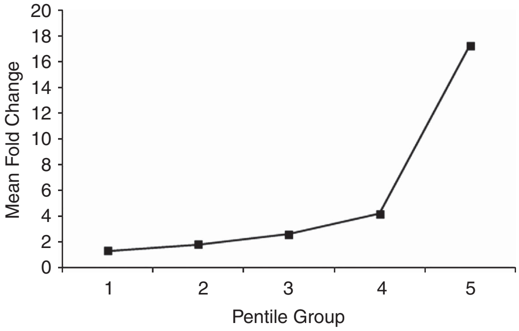


Figure 2 Mean values of antibody response of vaccinees belonging to the five pentile groups.

DISCUSSION

WHO has recommended that cholera vaccination be provided in many geographical regions.[2] Protective immune response to *V. cholerae* that colonizes the intestinal mucosal surface without invasion of enterocytes[45] is believed to reside at the mucosal surface.[44] Orally administered vaccines maximize the intestinal secretory AR, and a long-lasting memory so that the primed intestinal immune system can rapidly respond to a subsequent exposure to antigen.[46] Serum vibriocidal activity, measured by a bactericidal assay requiring the fixation of complement by antibody that is bound specifically to vibrios, is used extensively as a marker of immunity to *V. cholerae*, but has limitations.[47,48] On the basis of

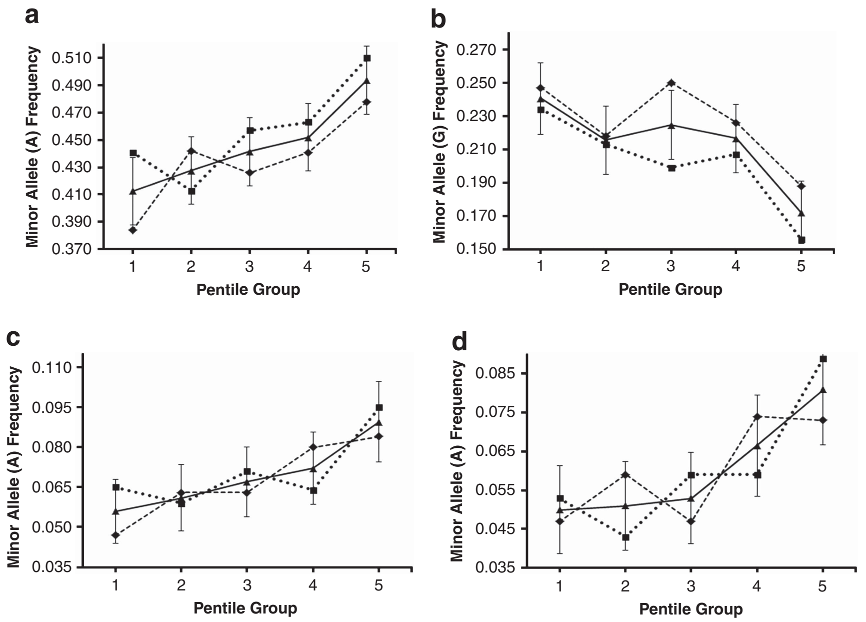


Figure 3 Minor allele frequencies at the significantly associated SNP loci in the association discovery (--◆--) and validation (---■---) half-samples, and in the pooled (—▲—) sample (±SD in the pooled sample) in PGs. (**a**) rs266087 (*CXCL12*), (**b**) rs598493 (*TNFAIP3*), (**c**) rs17180481 (*MARCO*) and (**d**) rs17180600 (*MARCO*).

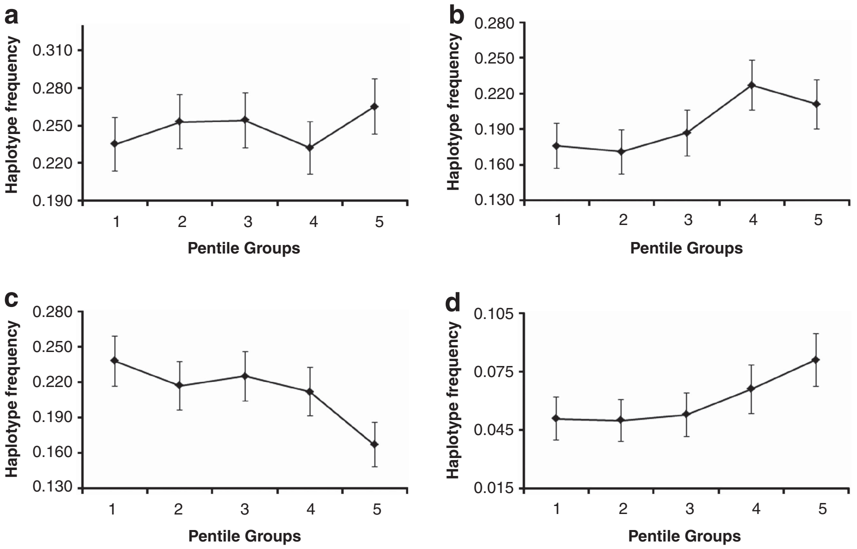


Figure 4 Frequencies of haplotypes containing the minor allele at the significantly associated SNP loci in the pooled sample PGs. (**a**) Haplotype GG**A**GCG of *CXCL12*, (**b**) Haplotype GA**A**GCG of *CXCL12*, (**c**) Haplotype GA**A**GC of *TNFAIP3*, (**d**) Haplotype **AA** of *MARCO*. Boldfaced and underlined alleles are those the minor alleles at the corresponding loci. (Another haplotype–GG**A**GC of *TNFAIP3*–had lower frequencies than those of the haplotype presented in this figure and was absent in three PGs.).

recent reccommendations[49] to measure LPS-response in multiple immunoglobulin isotypes and results[50,51] which indicate that levels of all immunoglobulin isotypes (IgG, IgM and IgA) determine the nature and extent of protection afforded by oral cholera vaccination, we have used the total anti-LPS AR and also the traditional vibriocidal assay.

Clinical trials of many oral vaccines have underscored the importance of understanding the reasons for variability in immune response to oral vaccines,[52] and the identification of genetic factors in the host linked to vaccine immunogenicity. It has been emphasized that such identification may result in adoption of 'inverse of personalized medicine' approaches to design vaccines to overcome barriers in nonresponding populations.[53] Susceptibility to cholera is significantly associated with a marker in the promoter region of *LPLUNC1* (rs11906665),[54] an innate immunity gene that likely has a role in modulating host inflammatory responses to *V. cholerae* infection.[55] We have discovered and validated that three genes–*MARCO*, *TNFAIP3* and *CXCL12*–has significant roles in modulating AR. However, *MARCO* was not significantly associated with increased vibriocidal titers. A single cell layer of the mucosal epithelium

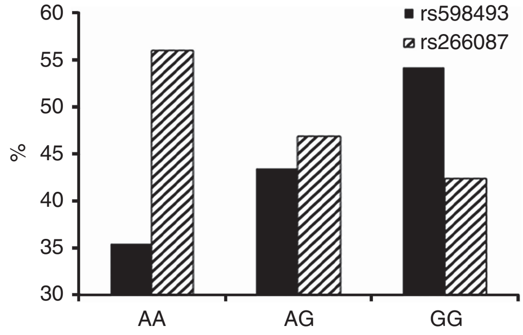


Figure 5 Percentages of vaccinees belonging to the three genotypes which elicited protective vibriocidal response (ie, seroconverted with fold-increase ≥4) upon vaccination for the two loci in *TNFAIP3* (rs598493) and *CXCL12* (rs266087).

essentially regulates intestinal homeostasis. This physical barrier is a prominent part of the innate immune system with damage-healing and infection-limiting properties. *CXCL12* is a homeostatic or constitutive chemokine that has a key role in healing by stimulating the activation of laminin-specific integrins.[56] *V. cholerae* infection produces strong inflammatory manifestations.[57] *CXCL12* is a neutrophil and lymphocyte chemoattractant that is upregulated in response to *V. cholerae* infection, resulting in increased recruitment of polymorphonuclear leukocytes to the site of infection.[58] The LPS in the cholera vaccine possibly provides signals that mimic those of the live bacterium, without actually causing disease. The TNF-induced protein 3 (*TNFAIP3*) is an ubiquitin-modifying enzyme and an essential negative regulator of inflammation, through the inhibition of NF-κB activation and prevention of TLR-mediated responses.[59] Tight junctions between intestinal epithelial cells mediate the permeability of the intestinal barrier. *TNFAIP3* has a role in promoting intestinal epithelial barrier integrity and in enabling maintenance of intestinal homeostasis through tight junction protein regulation;[60] possible requirements for adequate response to the whole-cell killed cholera vaccine used in this study. *TNFAIP3* has also been implicated in genome-wide studies of susceptibility of several autoimmune disorders.[61] *MARCO* is a type II transmembrane protein of the class A scavenger receptor family.[62] LPS, the major component of the vaccine used in this study, is a very potent activator of innate immune responses. LPS stimulation upregulates surface expression of *MARCO* in a dose- and time-dependent fashion.[63] *MARCO* is upregulated in bacterial infections in macrophages of most tissues; variants in this gene have crucial roles in bacterial binding.[64] No SNP in *MARCO* was found to be associated with high vibriocidal titer, possibly because different immunoglobulin isotypes are associated with anti-LPS and vibriocidal responses.[65]

The inferences of this study carried out in a cholera-endemic area, where many vaccinees possibly have prior exposure to the pathogen and consequently have high-baseline antibody levels (Table 2), may require validation in a non-endemic area. We, however, emphasize that the vaccine-response studies are more relevant and meaningful from a public-health viewpoint if conducted in endemic areas. In individuals with prior exposure to *V. cholerae*, the administered vaccine possibly acted as a booster, implying that individuals with high-baseline antibody levels are expected to show a low response (fold-change) to the vaccine, while individuals with low-baseline antibody levels are expected to show a high response. Although this feature was weakly observed (Table 2), the Day 0 and Day 28 antibody levels were significantly ($P<0.0001$) positively correlated ($r=0.449$). In this study we have used quantitative level of AR. Therefore, genomic associations with antibody fold-change detected in this study are unlikely to be affected by baseline antibody levels. However, it may be tricky is to relate the association between the level of AR and a SNP in a gene of the immune system with the level of immunological protection. In an endemic area, individuals with the lowest AR may be the best protected, while in a non-endemic area the reverse may be true. Although these issues pertaining to long-term protection remain unresolved, the genes that have shown significant association with AR are all biologically relevant. Functional studies on the markers found to be associated with AR need to be carried out, which together with the findings on *LPLUNC1*,[54,55] may provide a deep mechanistic understanding of the mode of action of the whole-cell killed oral cholera vaccine.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

Financial support for this work was provided by U.S. National Institute of Allergy and Infectious Diseases, National Institutes of Health Contract HHSN200400067C. The vaccine was used in this study with approval of the Drug Controller General of India, and the entire study was approved by the Health Ministry Monitoring Committee, Government of India; we are grateful to them for the approvals. We are grateful to members of CpG and TCGA for their expert help in the conduct of this project.

1 World Health Organization. Cholera: global surveillance summary, 2008. *Weekly Epid Rec* 2009; **84**: 309–324.
2 World Health Organization. Meeting of the Strategic Advisory Group of Experts on immunization October 2009–conclusions and recommendations. *Weekly Epid Rec* 2009; **50**: 526–528.
3 Date KA, Vicari A, Hyde TB *et al*: Considerations for oral cholera vaccine use during outbreak after earthquake in Haiti 2010-2011. *Emerg Infect Dis* 2011; **17**: 2105–2112.
4 Tappero JW, Tauxe-Emer RV: Lessons learned during public health response to cholera epidemic in haiti and the dominican republic. *Emerg Infect Dis* 2011; **17**: 2087–2093.
5 Institute of Medicine: *New Vaccine Development Establishing Priorities Volume II Diseases of Importance in Developing Countries*. Washington DC: National Academy Press, 1986.
6 World Health Organization: Cholera, 2010. *Weekly Epid Rec* 2011; **86**: 325–340.
7 Ryan ET: The cholera pandemic still with us after half a century Time to rethink. *PLoS Negl Trop Dis* 2011; **5**: e1003.
8 Nandy RK, Sengupta TK, Mukhopadhyay S, Ghosh AC: A comparative study of the properties of Vibrio cholera O139 O1 and other non-O1 strains. *J Med Microbiol* 1995; **42**: 251–257.
9 Levine MM, Black RE, Clements ML, Cisneros L, Nalin DR, Young CR: Duration of infection-derived immunity to cholera. *J Infect Dis* 1981; **143**: 818–820.
10 World Health Organization. Global Task Force on Cholera Control. Cholera Vaccines A new public health tool? *Report of a WHO meeting 10–11 December 2002 Geneva Switzerland World Health Organization* 2002, pp 14.
11 Sinclair D, Abba K, Zaman K, Qadri F, Graves PM: Oral vaccines for preventing cholera. *Cochrane Database Syst Rev* 2011; Issue 3 Art No CD008603.
12 Provenzano D, Kovác P, Wade WF: The ABCs (Antibody B cells and Carbohydrate epitopes) of cholera immunity considerations for an improved vaccine. *Microbiol Immunol* 2006; **50**: 899–927.
13 Clemens JD, Sack DA, Harris JR *et al*: ABO blood groups and cholera new observations on specificity of risk and modification of vaccine efficacy. *J Infect Dis* 1989; **159**: 770–773.
14 Glass RI, Holmgren J, Haley CE *et al*: Predisposition for cholera of individuals with O blood groupPossible evolutionary significance. *Am J Epidemiol* 1985; **121**: 791–796.
15 Harris JB, Khan AI, LaRocque RC *et al*: Blood group immunity and risk of infection with Vibrio cholerae in an area of endemicity. *Infect Immun* 2005; **73**: 7422–7427.
16 Lagos R, Avendaño A, Prado V *et al*: Attenuated live cholera vaccine strain CVD 103-HgR elicits significantly higher serum vibriocidal antibody titres in persons of blood group O. *Infect Immun* 1995; **63**: 707–709.
17 Ramamurthy T, Wagener D, Chowdhury G, Majumder PP: A large study on immunological response to a whole-cell killed oral cholera vaccine reveals that there are

significant geographical differences in response and that O blood group individuals do not elicit a higher response. *Clin Vaccine Immunol* 2010; **17**: 1232–1237.

18 Poland GA: Variability in immune response to pathogens using measles vaccine to probe immunogenetic determinants of response. *Am J Hum Genet* 1998; **62**: 215–220.

19 Poland GA, Jacobson RM, Colbourne SA *et al*: Measles antibody seroprevalence rates among immunized Inuit Innu and Caucasian subjects. *Vaccine* 1999; **17**: 1525–1531.

20 Milich DR, Leroux-Roels GG: Immunogenetics of the response to HBsAg vaccination. *Autoimmun Rev* 2003; **2**: 248–257.

21 Zuckerman JN: Nonresponse to hepatitis B vaccines and the kinetics of anti-HBs production. *J Med Virol* 1996; **50**: 283–288.

22 Gupta D, Faridi MM, Aggarwal A, Kaur I: Seroprevalence of anti-Vi antibodies and immunogenicity of Typhim Vi vaccine in children. *Hum Vaccine* 2008; **4**: 305–308.

23 Sur D, Ochiai RL, Bhattacharya SK *et al*: A cluster-randomized effectiveness trial of Vi typhoid vaccine in India. *N Engl J Med* 2009; **361**: 403–405.

24 Kimman TG, Vandebriel RJ, Hoebee B: Genetic variation in the response to vaccination. *Community Genet* 2007; **10**: 201–217.

25 Kruskall MS, Alper CA, Awdeh Z, Yunis EJ, Marcus-Bagley D: The immune response to hepatitis B vaccine in humans inheritance patterns in families. *J Exp Med* 1992; **175**: 495–502.

26 Gelder CM, Lambkin R, Hart KW *et al*: Associations between human leukocyte antigens and nonresponsiveness to influenza vaccine. *J Infect Dis* 2002; **185**: 114–117.

27 Poland GA, Ovsyannikova IG, Jacobson RM: Immunogenetics of seasonal influenza vaccine response. *Vaccine* 2008; **26**: D35–D40.

28 Poland GA, Ovsyannikova IG, Jacobson RM: Vaccine immunogenetics bedside to bench to population. *Vaccine* 2008; **26**: 6183–6188.

29 van Eden W, de Bries RRP, Stanford JL, Rook GAW: HLADR3 associated genetic control of response to multiple skin tests with new tuberculins. *Clin Exp Immunol* 1983; **52**: 287–292.

30 Wang C, Tang J, Song W, Lobashevsky E, Wilson CM, Kaslow RA: HLA and cytokine gene polymorphisms are independently associated with responses to hepatitis B vaccination. *Hepatology* 2004; **39**: 978–988.

31 Anh DD, Canh do G, Lopez AL *et al*: Safety and immunogenicity of a reformulated Vietnamese bivalent killed whole-cell oral cholera vaccine in adults. *Vaccine* 2007; **25**: 1149–1155.

32 Pulendran B, Kumar P, Cutler CW, Mohamadzadeh M, Van Dyke T, Banchereau J: Lipopolysaccharides from distinct pathogens induce different classes of immune responses in vivo. *J Immunol* 2001; **167**: 5067–5076.

33 Jin P, Wang E: Polymorphism in clinical immunology—from HLA typing to immunogenetic profiling. *J Transl Med* 2003; **1**: 8.

34 Majumder PP, Staats HF, Sarkar-Roy N *et al*: Genetic determinants of immune-response to a polysaccharide vaccine for typhoid. *HUGO J* 2009; **3**: 17–30.

35 Ovsyannikova IG, Haralambieva IH, Vierkant RA, Pankratz VS, Jacobson RM, Poland GA: The role of polymorphisms in Toll-like receptors and their associated intracellular signaling genes in measles vaccine immunity. *Hum Genet* 2011; **130**: 547–561.

36 Benenson AS, Saad A, Mosley WH: Serological studies in cholera 2 The vibriocidal antibody response of cholera patients determined by a microtechnique. *Bull World Health Organ* 1968; **38**: 277–285.

37 Sarkar-Roy N, Mondal D, Bhattacharya P, Majumder P: A novel statistical algorithm for enhancing the utility of HapMap data to design genomic association studies in non-HapMap populations. *Int J Data Mining Bioinf* 2011; **5**: 706–716.

38 Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D: Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 2006; **38**: 904–909.

39 Benjamini Y, Hochberg Y: Controlling the false discovery rate a practical and powerful approach to multiple testing. *J Roy Statist Soc Ser B Methodological* 1995; **57**: 289–300.

40 Conneely KN, Boehnke M: So many correlated tests so little time! Rapid adjustment of P values for multiple correlated tests. *Am J Hum Genet* 2007; **81**: 1158–1168.

41 Han B, Kang HM, Eskin E: Rapid and accurate multiple testing correction and power estimation for millions of correlated markers. *PLoS Genet* 2009; **5**: e1000456.

42 Stephens M, Smith N, Donnelly P: A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet* 2001; **68**: 978–989.

43 Eko FO, Schukovskaya T, Lotzmanova EY *et al*: Evaluation of the protective efficacy of Vibrio cholerae ghost VCG; candidate vaccines in rabbits. *Vaccine* 2003; **21**: 3663–3674.

44 Tarique AA, Kalsy A, Arifuzzaman M *et al*: Transcutaneous immunization with a Vibrio cholerae O1 Ogawa synthetic hexasaccharide conjugate following oral whole-cell cholera vaccination boosts vibriocidal responses and induces protective immunity in mice. *Clin Vaccine Immunol* 2012; **19**: 594–602.

45 Sack DA, Sack RB, Nair GB, Siddique AK: Cholera. *Lancet* 2004; **363**: 223–233.

46 Kaper JB, Morris JG, Levine MM: Cholera. *Clin Microbiol Rev* 1995; **8**: 48–86.

47 Clemens JD, van Loon F, Sack DA *et al*: Field trial of oral cholera vaccines in Bangladesh serum vibriocidal and antitoxic antibodies as markers of the risk of cholera. *J Infect Dis* 1991; **163**: 1235–1242.

48 Saha D, LaRocque RC, Khan AI *et al*: Incomplete correlation of serum vibriocidal antibody titer with protection from Vibrio cholerae infection in urban Bangladesh. *J Infect Dis* 2004; **189**: 2318–2322.

49 Losonsky GA, Yunyongying J, Lim V *et al*: Factors influencing secondary vibriocidal immune responses Relevance for understanding immunity to cholera. *Infect Immun* 1996; **64**: 10–15.

50 Alam MM, Riyadh MA, Fatema K *et al*: Antigen-specific memory B-cell responses in Bangladeshi adults after one- or two-dose oral killed cholera vaccination and comparison with responses in patients with naturally. *Clin Vaccine Immunol* 2011; **18**: 844–850.

51 Leung DT, Rahman MA, Mohasin M *et al*: Memory B Cell and Other Immune Responses in Children Receiving Two Doses of an Oral Killed Cholera Vaccine Compared to Responses following Natural Cholera Infection in Bangladesh. *Clin Vaccine Immunol* 2012; **19**: 690–698.

52 Levine MM: Immunogenicity and efficacy of oral vaccines in developing countries lessons from a live cholera vaccine. *BMC Biol* 2010; **8**: 129.

53 Serazin AC, Shackelton LA, Wilson C, Bhan MK: Improving the performance of enteric vaccines in the developing world. *Nat Immunol* 2010; **11**: 769–773.

54 LaRocque RC, Sabeti P, Duggal P *et al*: A variant in long palate lung and nasal epithelium clone 1 is associated with cholera in a Bangladeshi population. *Genes Immun* 2009; **10**: 267–272.

55 Shin OS, Uddin T, Citorik R *et al*: LPLUNC1 modulates innate immune responses to Vibrio cholera. *J Infect Dis* 2011; **204**: 1349–1357.

56 Agle KA, Vongsa RA, Dwinell MB: Chemokine stimulation promotes enterocyte migration through laminin-specific integrins. *Am J Physiol Gastrointest Liver Physiol* 2011; **301**: G968–G980.

57 Ou G, Rompikuntal PK, Bitar A *et al*: *Vibrio cholerae* cytolysin causes an inflammatory response in human intestinal epithelial cells that is modulated by the prtv protease. *PLoS ONE* 2009; **4**: e7806.

58 Poggi A, Carosio R, Fenoglio D *et al*: Migration of Vδ1 and Vδ2T cells in response to CXCR3 and CXCR4 ligands in healthy donors and HIV-1–infected patients competition by HIV-1 Tat. *Blood* 2003; **103**: 2205–2213.

59 Jennner RG, Young RA: Insights into host responses against pathogens from transcriptional profiling. *Nat Rev Microbiol* 2005; **3**: 281–294.

60 Kolodziej LE, Lodolce JP, Chang JE *et al*: TNFAIP2 maintains intestinal barrier function and supports epithelial cell tight junctions. *PLoS One* 2011; **6**: e26352.

61 Fung EY, Smyth DJ, Howson JM *et al*: Analysis of 17 autoimmune disease-associated variants in type 1 diabetes identifies 6q23/TNFAIP3 as a susceptibility locus. *Genes Immunity* 2009; **10**: 188–191.

62 Kraal G, van der Laan LJW, Elomaa O, Tryggvason K: The macrophage receptor MARCO. *Microbes Infect* 2000; **2**: 313–316.

63 Van der Laan LJW, Dopp EA, Haworth R *et al*: Regulation and functional involvement of macrophage scavenger receptor MARCO in clearance of bacteria in vivo. *J Immunol* 1999; **162**: 939–947.

64 Sankala M, Brännström A, Schulthess T *et al*: Characterization of Recombinant Soluble Macrophage Scavenger Receptor MARCO. *J Biol Chem* 2002; **277**: 33378–33385.

65 Alam MM, Riyadh MA, Fatema K *et al*: Antigen-specific memory B-cell responses in bangladeshi adults after one- or two-dose oral killed cholera vaccination and comparison with responses in patients with naturally acquired cholera. *Clin Vac Immuol* 2011; **18**: 844–850.

Supplementary Information accompanies the paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)

View publication stats

frontiers in Veterinary Science

ORIGINAL RESEARCH
published: 29 September 2020
doi: 10.3389/fvets.2020.547891

Genomic Epidemiology of *Salmonella* Infantis in Ecuador: From Poultry Farms to Human Infections

Lorena Mejía[1,2], José Luis Medina[3], Rosa Bayas[1], Carolina Satan Salazar[4], Fernando Villavicencio[4], Sonia Zapata[1], Jorge Matheu[5], Jaap A. Wagenaar[6,7], Fernando González-Candelas[2,8,9] and Christian Vinueza-Burgos[3]**

[1] Instituto de Microbiología, Colegio de Ciencias Biológicas y Ambientales, Universidad San Francisco de Quito, Quito, Ecuador, [2] Institute for Integrative Systems Biology, University of Valencia, Valencia, Spain, [3] Unidad de Investigación de Enfermedades Transmitidas por Alimentos y Resistencia a los Antimicrobianos (UNIETAR), Facultad de Medicina Veterinaria, Universidad Central del Ecuador, Quito, Ecuador, [4] Centro de Referencia Nacional de Resistencia a los Antimicrobianos, Instituto Nacional de Investigación en Salud Pública "Leopoldo Izquieta Pérez", Quito, Ecuador, [5] Department of Food Safety and Zoonoses, World Health Organization (WHO), Geneva, Switzerland, [6] Department of Infectious Diseases and Immunology, Faculty of Veterinary Medicine, Utrecht University, Utrecht, Netherlands, [7] Wageningen Bioveterinary Research (WBVR), Lelystad, Netherlands, [8] Joint Research Unit "Infection and Public Health" FISABIO-University of Valencia, Valencia, Spain, [9] CIBER (Centro de Investigación Biomédica en Red) in Epidemiology and Public Health, Valencia, Spain

OPEN ACCESS

Edited by:
Gabriel Gutkind,
University of Buenos Aires, Argentina

Reviewed by:
Pablo Chacana,
International Trademark Association,
United States
Nora Lía Padola,
National University of Central Buenos
Aires, Argentina

***Correspondence:**
Fernando González-Candelas
Fernando.Gonzalez@uv.es
Christian Vinueza-Burgos
cvinueza@uce.edu.ec

Specialty section:
This article was submitted to
Veterinary Epidemiology and
Economics,
a section of the journal
Frontiers in Veterinary Science

Received: *01 April 2020*
Accepted: *20 August 2020*
Published: *29 September 2020*

Citation:
Mejía L, Medina JL, Bayas R, Salazar CS, Villavicencio F, Zapata S, Matheu J, Wagenaar JA, González-Candelas F and Vinueza-Burgos C (2020) Genomic Epidemiology of Salmonella Infantis in Ecuador: From Poultry Farms to Human Infections.
Front. Vet. Sci. 7:547891.
doi: 10.3389/fvets.2020.547891

Salmonella enterica is one of the most important foodborne pathogens around the world. In the last years, *S. enterica* serovar Infantis has become an important emerging pathogen in many countries, often as multidrug resistant clones. To understand the importance of *S. enterica* in the broiler industry in Ecuador, we performed a study based on phenotypic and WGS data of isolates from poultry farms, chicken carcasses and humans. We showed a high prevalence of *S. enterica* in poultry farms (41.4%) and chicken carcasses (55.5%), but a low prevalence (1.98%) in human samples. *S.* Infantis was shown to be the most prevalent serovar with a 98.2, 97.8, and 50% in farms, foods, and humans, respectively, presenting multidrug resistant patterns. All sequenced *S.* Infantis isolates belonged to ST32. For the first time, a pESI-related megaplasmid was identified in Ecuadorian samples. This plasmid contains genes of antimicrobial resistance, virulence factors, and environmental stress tolerance. Genomic analysis showed a low divergence of *S.* Infantis strains in the three analyzed components. The results from this study provide important information about genetic elements that may help understand the molecular epidemiology of *S.* Infantis in Ecuador.

Keywords: *Salmonella* Infantis, ST32, broiler, WGS, Ecuador, megaplasmid, multidrug resistance (MDR)

INTRODUCTION

Foodborne infections caused by *Salmonella enterica* are of primary importance worldwide. The WHO estimates that *Salmonella* causes more than 153 million illnesses, 120,281 deaths, and 8.27 million disability-adjusted life years annually (1). As foodstuffs can be contaminated in several parts of the food chain, a "from farm to table" approach is necessary to understand the epidemiology of *Salmonella*. Although *Salmonella* can contaminate vegetables, food-producing animals, especially poultry, are considered important sources for human infections (2, 3). In Latin American countries, poultry is one of the main sources of protein of animal origin. This is the case of Ecuador, where poultry meat is the most consumed commodity with a yearly per capita consumption of 30.4 Kg (4).

Previous investigations in Latin America showed that *Salmonella enterica* serovar Infantis is an increasingly important serotype on poultry farms (5–9). Moreover, this serotype has also been reported to cause infections in local inhabitants and travelers that have visited Latin American countries (10–12). However, no genomic data considering isolates of *Salmonella* originated from animals, foodstuff, and humans have been released in Latin America so far.

Antimicrobials are commonly used in poultry production as both therapeutics and growth promoters. However, even when antimicrobials are used under technical criteria, they can select resistant strains of *Salmonella* that pose a public health problem when reaching consumers (13). This is of special concern in developing countries where the misuse of antimicrobials and lack of control is an issue to be addressed.

There is a wide diversity of virulence factors that are essential for pathogenicity of *Salmonella* in host cells. Among these factors, fimbriae, flagella, plasmids, pathogenicity islands, toxins, and secretion systems are the more frequently associated to pathogenic strains of *Salmonella* (14).

This research was aimed at describing by phenotyping methods and whole-genome sequencing, the antimicrobial resistance (AMR) characteristics and genetic profiles of *Salmonella* isolates obtained from broiler farms, broiler carcasses and humans in Quito—Ecuador.

MATERIALS AND METHODS

Study Design and Sampling

Samples for *Salmonella* isolation were collected weekly from November 2017 to November 2018 following the guidelines of the document: "Integrated Surveillance of Antimicrobial Resistance in Foodborne Bacteria" by the World Health Organization (15).

Poultry Farms: 133 flocks from 69 farms were sampled during the study period. For every sampled flock, 25 caeca from 25 chicken were randomly collected at the slaughterhouse level and transported to the laboratory in an ice-box within the next 2 h. Caecal samples are recommended by the WHO because they provide a higher recovery of isolates and better represent contamination of individual animals at the farm level (15). At the laboratory, a sample pool of 25 g was obtained for bacteriological isolation as previously described (8).

Chicken carcasses: 335 carcasses were collected in three kinds of markets as follows: 125 samples from supermarkets, 126 samples from small shops, and 84 samples from open markets. Sampling of chicken carcasses was performed alternately between the north and south of the city. Each carcass was collected in its original bag and transported to the laboratory in an ice-box within the next 2 h. At the laboratory, 25 g of breast skin of every carcass were aseptically collected for bacteriological analysis.

Human stool samples: 302 samples were evenly collected in two health care centers located in the urban periphery of Quito (Guamani health care center at the south and Calderon health care center at the north) from patients with two or more episodes of diarrhea or vomiting in the last 24 h. Human stool samples were transported to the laboratory in an ice-box within the next 2 h. Approximately 25 g of feces were collected for bacteriological analysis.

According to national legislation, ethics approval was not required for poultry farms and chicken carcasses sampling since no animals were sacrificed during this study. For the human component, the project was approved by the bioethics committee from the National Institute of Public Health "Leopoldo Izquieta Pérez" (Protocol ID:CEISH-INSPI-005). The participants were informed about the objective of the study and all volunteers provided a written consent. All personal information was anonymized.

Isolation and Identification of *Salmonella enterica*

Salmonella isolation was performed by a method based in the ISO 6579-1:2007 protocol. Briefly, 225 mL of Buffered Peptone Water (BPW; Difco, BD, Sparks, MD) was added to every sample, homogenized by hand for 1 min and incubated at 37°C for 20 h. Then, 100 μL of each enrichment was inoculated onto Modified Semi-solid Rappaport-Vassiliadis agar (MSRV; Oxoid, Basingstoke, UK) in three equidistant points and incubated at 42°C for 24 h. Afterwards, plates were examined for the presence of a white halo around of at least one inoculation point. A loopful taken from the edge of the white halo was streaked on a Xylose Lysine Deoxycholate agar (XLD, Difco) and incubated at 37°C for 24 h. After incubation, one suspect colony of *Salmonella* was biochemically confirmed by Triple Sugar Iron (TSI, Difco, BD), Iron Lysine (LIA, BBL, BD), Urea (BBL, BD), and Sulfur Indole Motility tests (SIM, BBL, BD). Isolated colonies were confirmed by PCR as previously described (16). The 95% confidence interval ($CI_{95\%}$) for the prevalence of *Salmonella* at each component was calculated.

Antimicrobial Susceptibility Testing

All *Salmonella* isolates were examined by the Kirby-Bauer disk diffusion method with the following antibiotics: sulfamethoxazole + trimethoprim (25 μg), gentamicin (10 μg), ciprofloxacin (5 μg), cefotaxime (30 μg), tetracycline (30 μg), streptomycin (10 μg), chloramphenicol (30 μg), cefoxitin (30 μg), amikacin (30 μg), nitrofurantoin (300 μg), azithromycin (15 μg), fosfomycin (200 μg), ertapenem (10 μg), amoxicillin + clavulanic acid (30 μg). *E. coli* ATCC 25922 strain was used as quality control. Results and methods were interpreted according to CLSI 2019 criteria considering all intermediate phenotypes as resistant for further analysis (17).

Detection of Extended Spectrum Beta-Lactamase (ESBL) Genes

Salmonella isolates that presented resistant phenotypes to beta-lactam antibiotics were further tested by PCR for the identification of ESBL genes. PCR conditions and primers were the ones described by Hasman et al. (18) for bla_{CTX-M}, Olesen et. al (19) for bla_{TEM}, Kruger et al. (20) for bla_{CMY} and Arlet et al. (21) for bla_{SHV}. Sub-families of bla_{CTX-M} genes were identified with PCR protocols described by Carattoli et al. (22) for $bla_{CTX-M-1}$, Jiang et al. (23) for $bla_{CTX-M-2}$, Hopkins et al. (24) for $bla_{CTX-M-8}$, Paauw et al. (25) for $bla_{CTX-M-9}$ and

Dierikx et al. (26) for $bla_{CTX-M-14}$. Amplification products were confirmed by gel electrophoresis using a 2% agarose gel. All PCR products were purified and sequenced at Macrogen Inc (Seul-South Korea). Obtained sequences were aligned against reference sequences with the online tool ResFinder v3.2 (27).

Whole Genome Sequencing

For whole genome sequencing (WGS), a selection of *Salmonella* isolates was made from the animal and food components. When selecting isolates from poultry farms the first positive sample of each farm was considered. For chicken carcasses, the first positive sample of every sampling week in each kind of market was selected. All non-*S.* Infantis isolates and all isolates from the human component were selected for WGS.

Genomic DNA was extracted using Invitrogen PureLink Genomic DNA Kit (Thermo Fisher Scientific, Walthman, MA, USA) following manufacturer's recommendations for Gram-negative bacterial cell lysates. DNA was quantified using Invitrogen Qubit 3.0 fluorometer (Thermo Fisher Scientific, Walthman, MA, USA), and sequenced with the Illumina NextSeq platform using Nextera XT Library Preparation Kit obtaining 150 × 2 bp paired-ends sequences (Illumina, San Diego, CA, USQ). Default parameters were used for all bioinformatic tools and programs unless otherwise specified. Reads were trimmed with Trimmomatic to remove ambiguous nucleotides and those with quality score values <20 (28). The programs Fastqc (29) and Multiqc (30) were used for quality assessment.

Serotype Identification

Salmonella serotypes were identified by PCR as described by Akiba et al. (16). Additionally, serotypes of isolates subjected to WGS were further confirmed by the analysis of their raw sequencing reads using the SeqSero pipeline (31).

MLST Analysis, Antimicrobial Resistance Genes, and Plasmid Detection

In order to identify MLST sequence types (ST), antimicrobial resistance genes and plasmid sequences, ARIBA (32) was used with PubMLST (33), ResFinder v3.2 (27), and PlasmidFinder 2.1 (34) databases, respectively. Phenotype resistance was compared with the presence of resistance genes found by WGS. Additionally, we performed a mapping against the megaplasmid p-F219 described by Vallejos-Sánchez et al. (35) using Burrows-Wheeler Aligner with BWA-MEM algorithm (36). BCFtools and vcfutils from SAMtools (37) were used to obtain the fastq files from SAM files, and the fasta sequences were transformed from fastq with Seqtk (38). The sequences were concatenated and a maximum likelihood phylogenetic tree was obtained with IQTREE2 software (39). As we obtained two well-defined clusters for *S.* Infantis, we annotated one representative of each plasmid cluster with Prokka (40) and performed an orthologous genes analysis (coverage of 90%, similarity on protein sequences of 80%) with Proteinortho5 (41). A manual comparison of all genes present in the plasmids was carried out. We performed the same analysis for non-Infantis isolates.

Megaplasmid Analysis

Two megaplasmids (pESI and p-F219) commonly associated to pathogenic and MDR strains of *S.* Infantis were analyzed in order to identify their relatedness. We used D-Genies (42) to obtain a dot plot of genome comparison, a genome alignment with progressiveMauve (43) in order to identify locally collinear blocks, and an ANI calculation (44) for computing average nucleotide identity in sequences shared by both plasmids.

Core Genome and Metadata Analysis

A Peruvian *S.* Infantis strain, FARPER-219 (35), and two Ecuadorian isolates (SRR4019589 and SRR4019602) analyzed by the US Centers for Disease Control and Prevention from two patients that developed salmonellosis after traveling to Ecuador (10) were added as references in the phylogenetic analysis of all *S.* Infantis isolates. From trimmed reads, Spades (45) was used to generate assemblies. Later, genome annotation was performed with Prokka (40). An orthologous genes analysis with the same conditions as for plasmid detection was performed with Proteinortho5 (41). The strict core genes, those present in all the isolates, were extracted with the Proteinortho tool: grab_proteins.pl. Mafft (46) and an in-house script were used for multiple alignment of every gene and subsequent concatenation in a single multiple alignment, respectively. The phylogenetic tree from the core genome alignment was obtained using IQTREE2 (39) with 1,000 bootstrap replicates. The metadata for sample origin, phenotypic antibiotic resistance patterns, and plasmid *in silico* detection was added to the final tree with iTol tools (47).

RESULTS

Salmonella Prevalence and Serotype Identification

Salmonella was present in 41.4% (55/133; $CI_{95\%}$:33–49.7), 55.5% (186/335; $CI_{95\%}$:50.2–60.8), and 1.98% (6/302; $CI_{95\%}$:0.4–3.6) in poultry farms, chicken carcasses and human stool samples, respectively. *S.* Infantis accounted for 98.2% ($n = 54$) of isolates from poultry farms, 97.8% ($n = 182$) of isolates from chicken carcasses, and one half ($n = 3$) of human sample isolates. Additionally, one isolate was typed as *S.* Enteritidis in broiler flocks; at the retail level one and three isolates were typed as *S.* Typhimurium and *S.* Enteritidis, respectively, while in the human stool samples two isolates were typed as *S.* Enteritidis and one isolate corresponded to monophasic *S.* Typhimurium 4,[5],12:i:- (**Supplementary Table 1**).

Antimicrobial Resistance

For *S.* Infantis isolates, antimicrobial resistance rates to nitrofurantoin, tetracycline, sulfamethoxazole + trimethoprim, streptomycin, gentamicin, cefotaxime, ciprofloxacin and chloramphenicol ranged from 64.8 to 100%. On the other hand, fosfomycin and azithromycin resistance rates were lower, ranging from 0 to 42.6%. Only one isolate from a stool sample presented phenotypic resistance to amikacin while none of the *Salmonella* isolates in this study was resistant to ertapenem (**Table 1**).

TABLE 1 | Number of *S*. Infantis isolates resistant to each tested antimicrobial.

Antimicrobial	**Number (%) of resistant isolates**		
	Poultry farms (farm)	**Chicken carcasses (food)**	**Stool samples (human)**
Nitrofurantoin	54 (100)	180 (99.4)	2 (66.7)
Tetracycline	54 (100)	176 (97.2)	3 (100)
Sulfamethoxazole + trimethoprim	44 (81.5)	158 (87.3)	1 (33.3)
Streptomycin	46 (85.2)	154 (85.1)	3 (100)
Gentamicin	45 (83.3)	155 (85.6)	2 (66.7)
Cefotaxime	51 (94.4)	150 (82.9)	1 (33.3)
Chloramphenicol	45 (83.3)	149 (82.3)	2 (66.7)
Ciprofloxacin	35 (64.8)	116 (64.1)	1 (33.3)
Fosfomycin	23 (42.6)	68 (37.6)	1 (33.3)
Azithromycin	10 (18.5)	31 (17.1)	0 (0)
Cefoxitin	7 (13)	11 (6.1)	0 (0)
Amoxicillin + clavulanic acid	7 (13)	7 (3.9)	0 (0)
Amikacin	0 (0)	0 (0)	1 (33.3)
Ertapenem	0 (0)	0 (0)	0 (0)

TABLE 2 | Antimicrobial resistance patterns of *S*. Enteritidis, *S*. Typhimurium, and monophasic *S*. Typhimurium 4,[5],12:i:-.

Resistant pattern	**No. Antimicrobial clases**	***S*. Enteritidis**	***S*. Typhimurium**	**Monophasic *S*. Typhimurium 4,[5],12:i:-**
SAQBTFN	7		1	
SAQTFNM	7	1		
SABTFNP	7	1		
SABTFN	6			1
QFM	3	1		
N	1	3*		

**1 isolate obtained from poultry farms and 2 isolates from human stool samples. Sulfonamide (S), aminoglycosides (A), quinolones (Q), Beta-lactams (B), tetracyclines (T), phenicol (F), nitrofuran (N), macrolides (M), Fosfomycin (P).*

Considering antimicrobial classes that were tested, *S*. Infantis isolates presented 43 antimicrobial resistance patterns. With the exception of one isolate from a stool sample, all isolates showed multidrug-resistant phenotypes. Importantly, 87 and 82% of isolates from poultry farms and chicken carcasses, respectively, presented resistance from 6 up to 9 classes of antimicrobials (**Supplementary Table 2**). One *S*. Infantis isolated from chicken carcasses could not be recuperated for this analysis.

Salmonella serotypes other than *S*. Infantis also presented multiresistant patterns, except for 3 *S*. Enteritidis isolates that were only resistant to one group of antimicrobials. For this set of isolates, every resistant pattern included isolates belonging to only one serotype (**Table 2**).

One isolate of *S*. Enteritidis, one of *S*. Typhimurium and one of Monophasic *S*. Typhimurium 4,[5],12:i:-; and 205 isolates of *S*. Infantis were identified as resistant to beta-lactam antibiotics.

Six *S*. Infantis isolates from chicken carcasses and one from poultry farms did not present any of the investigated ESBL genes. All other *S*. Infantis and one *S*. Enteritidis isolated from a carcass presented the $bla_{\mathrm{CTX-M-65}}$ gene.

Genomic Analysis

For WGS analysis, 144 isolates (40 from the animal component, 98 from the food component and six from the human component) were selected. Raw sequence data is available under bioproject PRJEB37560. The sequences from three samples were not enough to perform genomic analysis. The obtained average number of reads per strain was 1,356,678 (range 247,022–14,106,025) and after the quality control steps, the average number was 1,266,242 (range 228,263–13,094,594) (**Supplementary File 1**). Average Phred Score was Q34.

MLST typing showed that all *S*. Infantis isolates ($n = 137$) belonged to ST32. The five *S*. Enteritidis isolates belonged to ST11. Additionally, the single isolates of *S*. Typhimurium and monophasic *S*. Typhimurium 4,[5],12:i:- belonged to ST19 and ST2379, respectively (**Supplementary Table 1**).

The strict core genome of all *S*. Infantis included in the analysis corresponded to 3,552 genes and spanned 3,161,448 bp, 1,414 of which were variable (SNPs). The alignment of the concatenated genes present in this core was used to obtain a maximum-likelihood tree using FARPER 219 as outgroup (**Figure 1**). This strain was chosen because it was isolated in Peru, a neighbor country to Ecuador. Despite its inclusion in ST32, FARPER-219 presented genetic divergence with the Ecuadorian strains. The two *Salmonella* genomes from the USA (SRR4019589, SRR4019602) grouped indistinctive with some of the genomes of this study. Notably, the analyzed strains did not group according to their sampling origin or their phenotypic resistance patterns.

Genes of antimicrobial resistance were also confirmed with WGS data (**Supplementary File 2**). For most of the antimicrobial classes (folate pathway inhibition, aminoglycoside, beta-lactams, tetracyclines, fosfomycin, and phenicol) correspondence rates were higher than 80%. However, no genes responsible for the phenotypic resistance to quinolones, nitrofurans, and macrolides were found in sequenced isolates (**Table 3**).

Regarding virulence genes, *sifA, sseL, pipB, sopD2,* and *srlP* (part of SPI 2) were found in the core genome of all *S*. Infantis strains. The *lpf* operon that encodes the long polar fimbrae (LPF), the *fim* gene cluster that encodes type 1 fimbriae, and the *csg* operons that encodes the Tafi fimbriae (Thin aggregative fimbriae) were also present in these genomes.

The presence of plasmids was confirmed *in silico* by PlasmidFinder. Only one *S*. Infantis isolate (U2449s) presented one plasmid determinant (IncX1 and IncX1_1) despite the multidrug resistance patterns found in our isolates (**Figure 1** and **Supplementary File 2**). To further analyze the low incidence of plasmids, we mapped the raw reads from our samples against the megaplasmid p-F219 described by Vallejos-Sánchez et al. (35). We found two p-F219-like plasmids. The first one,

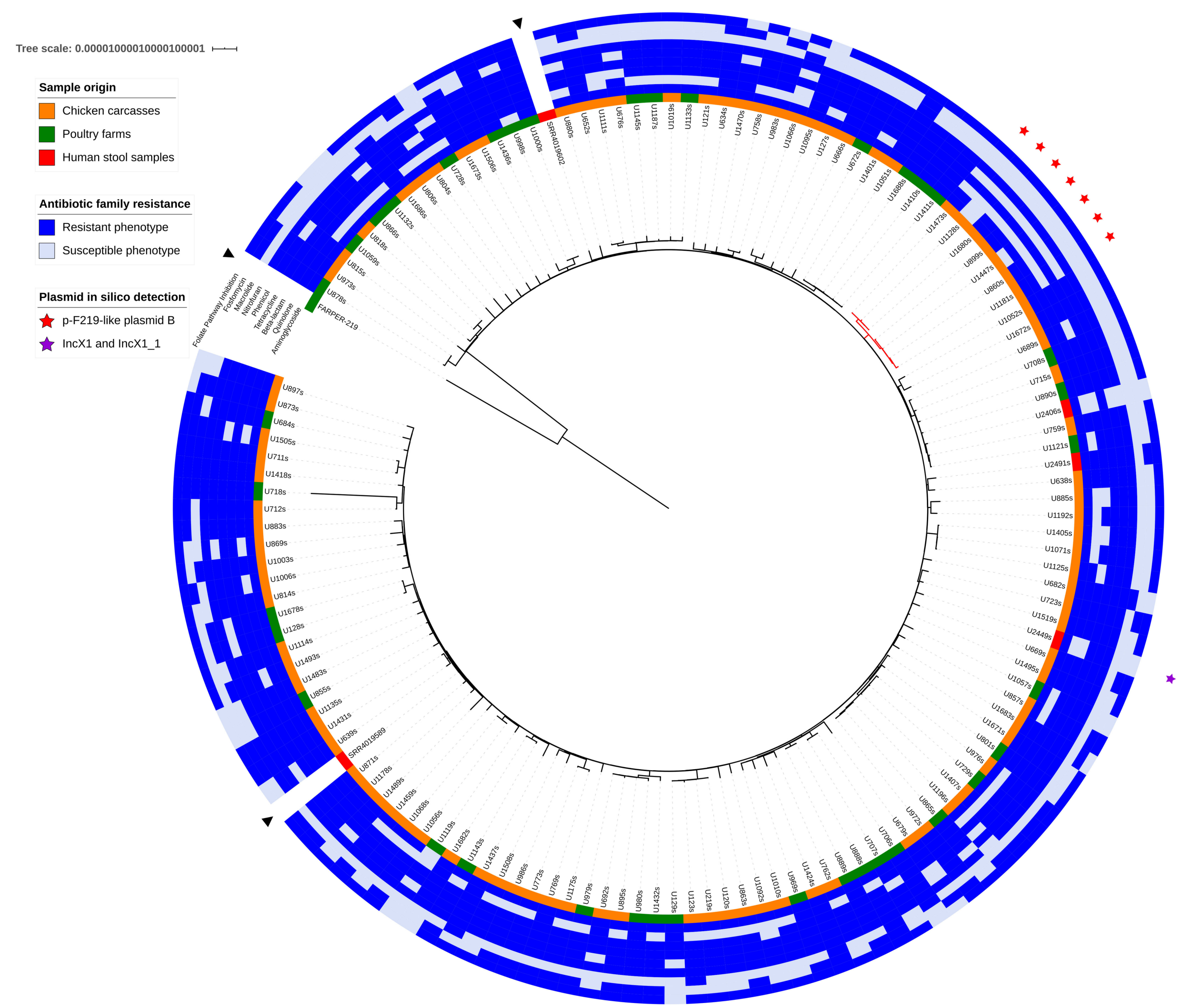

Tree scale: 0.00001000010000100001
Sample origin
🟧 Chicken carcasses
🟩 Poultry farms
🟥 Human stool samples
Antibiotic family resistance
🟦 Resistant phenotype
⬜ Susceptible phenotype
Plasmid in silico detection
★ p-F219-like plasmid B
★ IncX1 and IncX1_1
Folate Pathway Inhibition
Fosfomycin
Macrolide
Nitrofuran
Phenicol
Tetracycline
Beta-lactam
Quinolone
Aminoglycoside
FARPER-219
U878s
U973s
U815s
U1059s
U818s
U866s
U1132s
U1686s
U806s
U804s
U728s
U1673s
U1506s
U1436s
U998s
U1000s
SRR4019602
U880s
U652s
U1111s
U676s
U1145s
U1187s
U1019s
U1133s
U121s
U634s
U1470s
U758s
U983s
U1066s
U1095s
U1127s
U666s
U672s
U1401s
U1051s
U1588s
U1410s
U1411s
U1473s
U1128s
U1680s
U899s
U1447s
U860s
U1181s
U1052s
U1672s
U689s
U708s
U715s
U890s
U2406s
U759s
U1121s
U2491s
U638s
U885s
U1192s
U1405s
U1071s
U1125s
U682s
U723s
U1519s
U2449s
U669s
U1495s
U1057s
U857s
U1683s
U1671s
U801s
U976s
U729s
U1407s
U1196s
U872s
U865s
U679s
U709s
U707s
U888s
U889s
U762s
U1424s
U969s
U1010s
U1092s
U863s
U120s
U219s
U123s
U129s
U1432s
U980s
U895s
U692s
U979s
U1175s
U769s
U773s
U986s
U1508s
U1437s
U1143s
U1682s
U1119s
U1056s
U1068s
U1459s
U1489s
U1178s
U871s
SRR4019589
U639s
U1431s
U1135s
U855s
U1483s
U1493s
U1114s
U128s
U1678s
U814s
U1006s
U1003s
U869s
U883s
U712s
U718s
U1418s
U711s
U1505s
U684s
U873s
U897s

FIGURE 1 | Maximum likelihood phylogenetic tree of core genome alignment of all 137 *S.* Infantis isolates based on 3,552 genes. Two *S.* Infantis genomes from Ecuador detected in USA (SRR4019589, SRR4019602) and a Peruvian strain (FARPER-219) were included in the analysis and are indicated with a black triangle. The origin of each sample is colored in red for human stool samples, in green for poultry farms isolates and in orange for chicken carcasses strains. The phenotypic resistance for nine antibiotic families is marked with a blue box. Strains with p-F219-like plasmid B are marked with a red star. The rest of the samples harbor the p-F219-like plasmid A. The purple star indicates the presence of IncX1 and IncX1_1 plasmids in one of the strains. Digital version of the phylogenetic tree is available with iTOL login LMejia at https://itol.embl.de/shared_projects.cgi.

denoted as plasmid A, was present in most of the strains. This plasmid contained 338 genes shared with the p-F219 megaplasmid. The second one, denoted as plasmid B, was present in the remaining seven strains. Plasmid B lacked 72 of the genes present in plasmid A and presented six exclusive genes (**Supplementary File 3**). The strains that presented plasmid B belonged to a monophyletic clade (denoted with a red star in **Figure 1**). These strains also share susceptibility to fosfomycins, macrolides, phenicols, and beta-lactams. *Salmonella* strains harboring plasmid B were isolated from chicken carcasses sampled during different weeks of the year, different parts of the city and different types of retail stores (data not shown).

New hypothetical proteins ($n = 147$) were found in both plasmids; 43 exclusively found in plasmid A and five in plasmid B (**Supplementary File 3**).

The comparison of the p-F219 plasmid with another megaplasmid, commonly found in pathogenic *S.* Infantis strains, pESI plasmid, showed that they share more than 79% of their sequences (>75% of identity) (**Supplementary Figure 1**). Besides, we noticed a large genomic inversion in plasmid p-F219 when compared with the pESI plasmid that is also observable in progressiveMauve genome alignment (**Supplementary Figure 2**). From the ANI calculation, 99.41% of identity was found in the orthologous genes present in both samples.

TABLE 3 | Comparison of phenotypic AMR with AMR genes obtained from WGS data.

Antibiotic family	Phenotype (%)[a]	Phenotype + AMR gene[b] (%)	No phenotype + AMR gene[c] (%)
Folate pathway inhibition	84.56	92.17	7.83
Aminoglycoside	97.06	100	3.03
Quinolone	58.09	1.27	0
Beta-lactams	84.56	94.87	4.27
Tetracycline	97.06	99.24	0.76
Phenicol	80.88	80.91	6.36
Nitrofuran	98.53	0	0
Macrolide	17.65	0	0
Fosfomycin	37.5	90.2	41.18

[a] *Rate of isolates with phenotypic resistance.*

[b] *Rate of isolates with phenotypic resistance that presented a resistance gene by WGS analysis.*

[c] *Rate of isolates without phenotypic resistance that presented a resistance gene by WGS analysis.*

In silico plasmid detection in non-Infantis isolates showed the presence of two plasmids in all *S.* Enteritidis strains, while monophasic *S.* Typhimurium 4,[5],12:i:- and *S.* Typhimurium presented 1 and 5 plasmids, respectively (**Supplementary Table 1**).

DISCUSSION

In the last decades, there has been a clear rise in the prevalence of multidrug resistant *Salmonella enterica* worldwide, especially of serovar Infantis (11, 48–50).

To better explain the epidemiology of *Salmonella* in Ecuador, we studied *S. enterica* isolated from poultry farms, chicken carcasses and human stool samples in Quito. The prevalence of *Salmonella* was high in poultry farms and chicken carcasses, similarly to other studies in the region (6–8, 51). Although this research did not look for *Salmonella* in earlier stages of the broiler production chain, previous studies in Ecuador have reported the importance of compound feed, 1-day-old chicks and broiler pens in the *Salmonella* exposure of broilers (7, 52). These results highlight the necessity to improve broilers production systems in Ecuador toward a better control of *Salmonella* in the food chain.

In this study, it was also seen that samples from different kinds of markets delivered similar rates of *Salmonella* isolates, denoting that the geographical distribution of retailers does not influence the presence of *Salmonella* in carcasses (**Supplementary Table 3**). On the other hand, all human isolates originated in the southern health care center. However, it must to be considered that the low *Salmonella* prevalence in human samples could be influenced by the fact that other pathogens might be the main causes of diarrhea. This fact has been studied in Ecuador were other viruses, parasites and bacteria are the main cause of gastroenteritis cases (53–56). Additionally, the state of health carriers should be considered when accessing the real prevalence of *Salmonella* in humans (57). These circumstances represent limitations of this study and should be considered in future research.

The predominance of *S.* Infantis in this study is in accordance with other reports in the world that show that this serotype is becoming an emergent pathogen. For example, in Europe, *S.* Infantis has been reported to be one of the most common serovars in poultry and in humans (58). The same tendency has been reported in the neighboring country of Peru, where *S.* Infantis is the most prevalent serotype in broilers (6). However, a wider variety of serotypes has been reported in other Latin American countries (51, 59, 60). This could be explained by the fact that the poultry industry of Peru and Ecuador have close commercial interactions which could determine a common epidemiology of this pathogen. Nevertheless, a recent publication from Chile reported that 24% of broiler meat samples ($n = 361$) were positive to the isolation of *S.* Infantis (61). This data highlights the necessity of more research in the field to better understand the epidemiology of *Salmonella* in the region.

Several studies have shown that control programs of targeted *Salmonella* serotypes could have favored the occurrence of other serotypes (2, 62). This kind of shifts might explain to some extent the low presence of *S.* Enteritidis and Typhimurium among our samples but further research is needed to prove this hypothesis.

Most *Salmonella* isolates in this study presented multidrug resistance (MDR) phenotypes. This issue is especially evident in *S.* Infantis, as already been seen in Ecuador (7, 8) and other countries of the region (63–65). Although in a lower extent, this feature has also been reported in Europe where high rates of MDR (up to 85%) are reported in *S.* Infantis isolated from poultry (66). The high levels of resistance in *Salmonella* isolates in Latin America could be related to the intensive use of antimicrobials in poultry production as prophylactics, therapeutics, and growth promotors (67, 68).

S. Infantis isolates belonged to ST32, that is among those more frequently identified within this serovar (33, 69). All Ecuadorian *S.* Infantis isolates showed a high genomic similarity with an apparently common origin. However, in order to verify that all in fact share a common ancestor, a larger analysis including isolates from countries around the world is necessary for identifying the origin of this serovar in Ecuador. The close similarity found between isolates from farms, animals and humans show that this pathogen may be responsible for human infections through the food chain.

We did not observe any clear clustering of *S.* Infantis isolates and antibiotic phenotypic resistance patterns, what makes sense since the phylogenetic tree was obtained from the core genome and most of the antimicrobial resistance determinants are expected to be part of the accessory genome.

The resistance genes analysis carried out here could not successfully explain resistance to macrolides, nitrofurans, and quinolones in our isolates. In fact, phenotypic patterns and genetic detection correlation differences have been described previously (70–72), but point mutations not considered in this study may explain these phenotypes. Additional studies are still needed in order to identify the genes or mutations responsible for antimicrobial resistance in *S.* Infantis.

We found two p-F219-like plasmids (named A and B) present in all the analyzed isolates. Our samples share a significant number of genes with the p-F219 plasmid. The similarity between both p-F219-like plasmids requires a recent common ancestor from which they may have evolved. These plasmids may provide fitness and pathogenic advantages to the strains since they contain several genes for antimicrobial resistance, fimbriae, transposases, and environmental stress tolerance (**Supplementary File 3**). Other *S.* Infantis isolates around the world have been shown to also harbor pESI-related megaplasmids, that appear to be the difference between non-pathogenic and pathogenic isolates since it also provides antimicrobial resistance, oxidative stress tolerance, pathogenicity traits, and mercury environmental tolerance (69, 70, 73, 74). Besides, pESI has shown more pathogenicity and increased intestinal inflammation in experimental mice infections when compared to the plasmid-free isolates (73).

Both plasmids (p-F219 and pESI) share regions of very close similarity with an extensive genomic inversion. Genes present in both plasmids share more than 99% average nucleotide identity suggesting that p-F219 is actually a pESI-like plasmid and the variants found in our samples may also be cataloged as such. A genomic comparison between *S.* Infantis strains showed that large plasmids from multiple isolates actually share a pESI backbone with some genetic plasticity to add different mobile genetic elements due to insertion sequences (75). As pESI-like plasmids confer a MDR phenotype but also several virulence factors and tolerance to environmental stress, their acquisition may have been involved for making *S.* Infantis a successful emerging pathogen worldwide.

The negative results from the unmapped reads from *S.* Enteritidis and *S.* Typhimurium isolates against the p-F219 plasmid suggest that this plasmid may not be widespread among other *Salmonella* serovars, but more isolates need to be analyzed to confirm this hypothesis.

Other virulence genes in *Salmonella* are essential for pathogenicity and infection. We looked for some of the more relevant virulence and invasion determinants in the core genome of *S.* Infantis as determined by ProteinOrtho. Some genes that are usually identified as part of the *Salmonella* pathogenicity island 2 (SPI-2) were found: *sifA, sseL, pipB, sopD2, srlP* (75). The long polar fimbrae (LfP), encoded by the *lpf* operon, was found as putative proteins in all isolates; it is believed that this protein is involved in adhesion and growth on the small intestine mucosa (76), evasion of the host immune system (77). Moreover, variations in these genes may also impact the host range of the bacteria (78). We also found the *fim* gene cluster that encodes type 1 fimbriae and it is involved in the initiation of biofilm formation (79). Another determinant found in our strains was the *csg* operon that encodes the Tafi fimbriae (Thin aggregative fimbriae). Tafi is responsible of adhesive activities and biofilm formation (76). The presence of these genes in *S.* Infantis may suggest a pathogenic character of this serotype that, together with the multidrug resistant profiles, represent a potential public health concern.

To the best of our knowledge this is the first report based on *Salmonella enterica* in Ecuador that uses phenotypic and WGS information to analyze the relatedness of strains isolated from poultry, food and human samples. Isolates from this study show multidrug resistance patterns highlighting the importance of a reduced and better usage of antimicrobials in intensive poultry farms settings. Presence of related megaplasmids together with the core genome high similarity may suggest the dissemination of *S.* Infantis through the food chain to humans. The data presented here has shown the importance of *Salmonella enterica* serovar Infantis as a foodborne pathogen in Ecuador and provide critical information about its clonality and circulating strains.

DATA AVAILABILITY STATEMENT

Sequence data are available under bioproject PRJEB37560.

ETHICS STATEMENT

According to national legislation, ethics approval was not required for poultry farms and chicken carcasses sampling since no animals were sacrificed during this study. For the human component, the project was approved by the bioethics committee from the National Institute of Public Health Leopoldo Izquieta Pérez (Protocol ID was CEISH-INSPI-005). The participants were informed about the study's objective and all volunteers provided a written consent. All personal information was anonymized.

AUTHOR CONTRIBUTIONS

JM, JW, FV, and CV-B conception and design of the study. JLM, CS, FV, and RB performed laboratory analysis. JLM and CV-B performed the statistical analysis. JLM, LM, and CV-B organized the databases. LM and FG-C performed the bioinformatic and genomic analysis. CV-B and LM wrote the first draft of the manuscript. LM, SZ, FG-C, and CV-B contributed to the manuscript revision. All authors contributed to the article and approved the submitted version.

FUNDING

This project was funded by the WHO Advisory Group on Integrated Survillance on AMR (AGISAR) grant AGISAR-JM/lg and also funded by projects BFU2017-89594R from MICIN (Spanish Government) and PROMETEO2016-0122 (Generalitat Valenciana, Spain). WGS was co-financed by the European Union through the Operational Program of European Regional Development Fund (ERDF) of Valencia Region (Spain) 2014–2020. LM was recipient of a predoctoral fellowship from Fundación Carolina and Universidad San Francisco de Quito.

ACKNOWLEDGMENTS

Our acknowledgments to Dr. Gina Watson and Dr. Aida Soto from the OPS/OMS Ecuadorian office for the support in the logistical management of this project. We also thank Mariuxi Japon and Paola Damian for their support with the laboratory work and isolation of *Salmonella*. Whole genome sequences were obtained at Servicio de Secuenciación y Bioinformática (FISABIO, Valencia). Action co-financed by the European Union through the Operational Program of European Regional Development Fund (ERDF) of Valencia Region (Spain) 2014–2020.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fvets.2020.547891/full#supplementary-material

REFERENCES

1. WHO. *WHO Estimates of the Global Burden of Foodborne Diseases. Foodborne Disease Burden Epidemiolgy Reference Group 2007–2015*. Geneva: World Heal Organization (2015). p. 1–15.
2. Antunes P, Mourão J, Campos J, Peixe L. Salmonellosis: the role of poultry meat. *Clin Microbiol Infect.* (2016) 22:110–21. doi: 10.1016/j.cmi.2015.12.004
3. Koutsoumanis K, Allende A, Alvarez-Ordóñez A, Bolton D, Bover-Cid S, Chemaly M, et al. Salmonella control in poultry flocks and its public health impact. *EFSA J.* (2019) 17:e05596. doi: 10.2903/j.efsa.2019.5596
4. CONAVE. *Estadísticas Del Sector Avícola*. (2020) Available online at: https://www.conave.org/informacion-sector-avicola-publico/ (accessed February 21, 2020).
5. Donado-Godoy P, Byrne BA, León M, Castellanos R, Vanegas C, Coral A, et al. Prevalence, resistance patterns, and risk factors for antimicrobial resistance in bacteria from retail chicken meat in Colombia. *J Food Prot.* (2015) 78:751–9. doi: 10.4315/0362-028X.JFP-14-349
6. Valderrama W, Pastor J, Mantilla Salazar J, Ortiz M. Estudio de prevalencia de serotipos de salmonella en granjas avícolas tecnificadas en el perú. Lima. (2014).
7. Vinueza-Burgos C, Baquero M, Medina J, De Zutter L. Occurrence, genotypes and antimicrobial susceptibility of salmonella collected from the broiler production chain within an integrated poultry company. *Int J Food Microbiol.* (2019) 299:1–7. doi: 10.1016/j.ijfoodmicro.2019.03.014
8. Vinueza-Burgos C, Cevallos M, Ron-Garrido L, Bertrand S, De Zutter L. Prevalence and diversity of salmonella serotypes in ecuadorian broilers at slaughter age. *PLoS ONE.* (2016) 11:e0159567. doi: 10.1371/journal.pone.0159567
9. Voss-Rech D, Vaz CSL, Alves L, Coldebella A, Leao JA, Rodrigues DP, et al. A temporal study of salmonella enterica serotypes from broiler farms in Brazil. *Poult Sci.* (2015) 94:433–41. doi: 10.3382/ps/peu081
10. Brown AC, Chen JC, Watkins LKF, Campbell D, Folster JP, Tate H, et al. CTX-M-65 extended-spectrum β-lactamase–producing salmonella enterica serotype infantis, united states1. *Emerg Infect Dis.* (2018) 24:2284–91. doi: 10.3201/eid2412.180500
11. Cartelle Gestal M, Zurita J, Paz y Mino A, Ortega-Paredes D, Alcocer I. Characterization of a small outbreak of salmonella enterica serovar infantis that harbour CTX-M-65 in ecuador. *Brazilian J Infect Dis.* (2016) 20:406–7. doi: 10.1016/j.bjid.2016.03.007
12. Tate H, Folster JP, Hsu C-H, Chen J, Hoffmann M, Li C, et al. Comparative analysis of extended-spectrum-β-lactamase CTX-M-65-producing salmonella enterica serovar infantis isolates from humans, food animals, and retail chickens in the united states. *Antimicrob Agents Chemother.* (2017) 61:1–11. doi: 10.1128/AAC.00488-17
13. McDermott PF, Zhao S, Tate H. Antimicrobial resistance in nontyphoidal salmonella. *Microbiol Spectr.* (2018) 6:780–90. doi: 10.1128/9781555819804.ch12
14. Ibarra JA, Steele-Mortimer O. Salmonella - the ultimate insider. Salmonella virulence factors that modulate intracellular survival. *Cell Microbiol.* (2009) 11:1579–86. doi: 10.1111/j.1462-5822.2009.01368.x
15. WHO. *Integrated Surveillance of Antimicrobial Resistance in Foodborne Bacteria*. (2017). Available online at: http://apps.who.int/iris/bitstream/10665/91778/1/9789241506311_eng.pdf (accessed July 25, 2020).
16. Akiba M, Kusumoto M, Iwata T. Rapid identification of salmonella enterica serovars, typhimurium, choleraesuis, infantis, hadar, enteritidis, dublin and gallinarum, by multiplex PCR. *J Microbiol Methods.* (2011) 85:9–15. doi: 10.1016/j.mimet.2011.02.002
17. CLSI (Clinical Laboratory Standards Institute). Performance standards for antimicrobial susceptibility testing; twenty-nine informational supplement. In: Weinstein MP, editor. *CLSI Document M100-S29*. 29th ed. Wayne: Clinical and Laboratory Standards Institute. (2019). p. 32–41.
18. Hasman H, Mevius D, Veldman K, Olesen I, Aarestrup FM. β-Lactamases among extended-spectrum beta-lactamase (ESBL)-resistant Salmonella from poultry, poultry products and human patients in The Netherlands. *J Antimicrob Chemother.* (2005) 56:115–21. doi: 10.1093/jac/dki190
19. Olesen I, Hasman H, Aarestrup FM. Prevalence of beta-lactamases among ampicillin-resistant Escherichia coli and salmonella isolated from food animals in Denmark. *Microb Drug Resist.* (2004) 10:334–40. doi: 10.1089/mdr.2004.10.334
20. Kruger T, Szabo D, Keddy KH, Deeley K, Marsh JW, Hujer AM, et al. Infections with nontyphoidal salmonella species producing TEM-63 or a novel TEM enzyme, TEM-131, in South Africa. *Antimicrob Agents Chemother.* (2004) 48:4263–70. doi: 10.1128/AAC.48.11.4263-4270.2004
21. Arlet G, Rouveau M, Philippon A. Substitution of alanine for aspartate at position 179 in the SHV-6 extended-spectrum beta-lactamase. *FEMS Microbiol Lett.* (1997) 152:163–7. doi: 10.1016/S0378-1097(97)00196-1
22. Carattoli A, García-Fernández A, Varesi P, Fortini D, Gerardi S, Penni A, et al. Molecular epidemiology of *Escherichia coli* producing extended-spectrum beta-lactamases isolated in Rome, Italy. *J Clin Microbiol.* (2008) 46:103–8. doi: 10.1128/JCM.01542-07
23. Jiang X, Zhang Z, Li M, Zhou D, Ruan F, Lu Y. Detection of extended-spectrum beta-lactamases in clinical isolates of pseudomonas aeruginosa. *Antimicrob Agents Chemother.* (2006) 50:2990–5. doi: 10.1128/AAC.01511-05
24. Hopkins KL, Batchelor MJ, Liebana E, Deheer-Graham AP, Threlfall EJ. Characterisation of CTX-M and AmpC genes in human isolates of *Escherichia coli* identified between 1995 and 2003 in England and Wales. *Int J Antimicrob Agents.* (2006) 28:180–92. doi: 10.1016/j.ijantimicag.2006.03.027
25. Paauw A, Fluit AC, Verhoef J, Leverstein-van Hall MA. Enterobacter cloacae outbreak and emergence of quinolone resistance gene in Dutch hospital. *Emerg Infect Dis.* (2006) 12:807–12. doi: 10.3201/eid1205.050910
26. Dierikx CM, van Duijkeren E, Schoormans AHW, van Essen-Zandbergen A, Veldman K, Kant A, et al. Occurrence and characteristics of extended-spectrum-β-lactamase- and AmpC-producing clinical isolates derived from companion animals and horses. *J Antimicrob Chemother.* (2012) 67:1368–74. doi: 10.1093/jac/dks049
27. Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, et al. Identification of acquired antimicrobial resistance genes. *J Antimicrob Chemother.* (2012) 67:2640–4. doi: 10.1093/jac/dks261
28. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics.* (2014) 30:2114–20. doi: 10.1093/bioinformatics/btu170
29. Wingett SW, Andrews S. FastQ screen: a tool for multi-genome mapping and quality control. *F1000Research.* (2018) 7:1338. doi: 10.12688/f1000research.15931.2
30. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics.* (2016) 32:3047–8. doi: 10.1093/bioinformatics/btw354
31. Zhang S, Yin Y, Jones MB, Zhang Z, Deatherage Kaiser BL, Dinsmore BA, et al. Salmonella serotype determination utilizing high-throughput genome sequencing data. *J Clin Microbiol.* (2015) 53:1685–92. doi: 10.1128/JCM.00323-15
32. Hunt M, Mather AE, Sánchez-Busó L, Page AJ, Parkhill J, Keane JA, et al. ARIBA: rapid antimicrobial resistance genotyping directly from sequencing reads. *Microb Genomics.* (2017) 3:e000131. doi: 10.1099/mgen.0.000131
33. Jolley KA, Bray JE, Maiden MCJ. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. *Wellcome Open Res.* (2018) 3:124. doi: 10.12688/wellcomeopenres.14826.1

34. Carattoli A, Zankari E, Garciá-Fernández A, Larsen MV, Lund O, Villa L, et al. *In silico* detection and typing of plasmids using plasmidfinder and plasmid multilocus sequence typing. *Antimicrob Agents Chemother.* (2014) 58:3895–903. doi: 10.1128/AAC.02412-14
35. Vallejos-Sánchez K, Tataje-Lavanda L, Villanueva-Pérez D, Bendezú J, Montalván Á, Zimic-Peralta M, et al. Whole-genome sequencing of a salmonella enterica subsp. *Enterica serovar* infantis strain isolated from broiler chicken in Peru. *Microbiol Resour Announc.* (2019) 8:e00826–19. doi: 10.1128/MRA.00826-19
36. Li H, Durbin R. Fast and accurate long-read alignment with burrows-wheeler transform. *Bioinformatics.* (2010) 26:589–95. doi: 10.1093/bioinformatics/btp698
37. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. 1000 genome project data processing subgroup. The sequence alignment/map format and SAMtools. *Bioinformatics.* (2009) 25:2078–9. doi: 10.1093/bioinformatics/btp352
38. Li H. *Seqtk: Toolkit for Processing Sequences in FASTA/Q Formats* Hinxton. (2012).
39. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol Biol Evol.* (2020) 37:1530–34. doi: 10.1101/849372
40. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics.* (2014) 30:2068–9. doi: 10.1093/bioinformatics/btu153
41. Lechner M, Findeiss S, Steiner L, Marz M, Stadler PF, Prohaska SJ. Proteinortho: detection of (co-)orthologs in large-scale analysis. *BMC Bioinformatics.* (2011) 12:124. doi: 10.1186/1471-2105-12-124
42. Cabanettes F, Klopp C. D-GENIES: dot plot large genomes in an interactive, efficient and simple way. *PeerJ.* (2018) 6:e4958. doi: 10.7717/peerj.4958
43. Darling AE, Mau B, Perna NT. Progressivemauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS ONE.* (2010) 5:e11147. doi: 10.1371/journal.pone.0011147
44. Yoon S-H, Ha S-M, Lim J, Kwon S, Chun J. A large-scale evaluation of algorithms to calculate average nucleotide identity. *Antonie Van Leeuwenhoek.* (2017) 110:1281–6. doi: 10.1007/s10482-017-0844-4
45. Nurk S, Bankevich A, Antipov D, Gurevich A, Korobeynikov A, Lapidus A, et al. Assembling genomes mini-metagenomes from highly chimeric reads. In: M Deng, R Jiang, F Sun, X Zhang, editors. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence Lecture Notes in Bioinformatics).* Berlin: Springer (2011). p. 158–170. doi: 10.1007/978-3-642-37195-0_13
46. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic Acids Res.* (2002) 30:3059–66. doi: 10.1093/nar/gkf436
47. Letunic I, Bork P. Interactive tree of life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* (2019) 47:W256–9. doi: 10.1093/nar/gkz239
48. Hendriksen RS, Vieira AR, Karlsmose S, Lo Fo Wong DMA, Jensen AB, Wegener HC, et al. Global monitoring of salmonella serovar distribution from the world health organization global foodborne infections network country data bank: results of quality assured laboratories from 2001 to 2007. *Foodborne Pathog Dis.* (2011) 8:887–900. doi: 10.1089/fpd.2010.0787
49. EFSA (European Food Safety Authority), ECDC (European Centre for Disease Prevention and Control). The European union summary report on trends and sources of zoonoses, zoonotic agents and food-borne outbreaks in 2012. *EFSA J.* (2014) 12:e5077. doi: 10.2903/j.efsa.2014.3547
50. Gymoese P, Kiil K, Torpdahl M, Østerlund MT, Sørensen G, Olsen JE, Nielsen EM, et al. WGS based study of the population structure of salmonella enterica serovar infantis. *BMC Genomics.* (2019) 20:870. doi: 10.1186/s12864-019-6260-6
51. Donado-Godoy P, Gardner I, Byrne BA, Leon M, Perez-Gutierrez E, Ovalle MV, et al. Prevalence, risk factors, and antimicrobial resistance profiles of salmonella from commercial broiler farms in two important poultry-producing regions of Colombia. *J Food Prot.* (2012) 75:874–83. doi: 10.4315/0362-028X.JFP-11-458
52. Villagomez S. *Aislamiento y serotipificación de salmonella enteritidis, typhimurium e infantis en carcasas de pollo destinadas para consumo humano en un camal industrializado de la provincia de pichincha* Quito: Universidad Central del Ecuador. (2015)
53. Vasco G, Trueba G, Atherton R, Calvopiña M, Cevallos W, Andrade T, et al. Identifying etiological agents causing diarrhea in low income ecuadorian communities. *Am J Trop Med Hyg.* (2014) 91:563–9. doi: 10.4269/ajtmh.13-0744
54. Naranjo A, Cedeño C, Teran E, Castello A, CASERO Research Team. Prevalence of VP4 and VP7 genotypes of human rotavirus in ecuadorian children with acute diarrhea. *J Med Virol.* (2008) 80:1106–11. doi: 10.1002/jmv.21181
55. Guderian RH, Ordóñez G, Bossano R. Acute diarrhea associated with a campylobacter and other pathogenic agents in Quito, Ecuador. *Boletín Of Sanit Panam.* (1987) 102:333–9.
56. Bhavnani D, Goldstick JE, Cevallos W, Trueba G, Eisenberg JNS. Synergistic effects between rotavirus and coinfecting pathogens on diarrheal disease: evidence from a community-based study in northwestern ecuador. *Am J Epidemiol.* (2012) 176:387–95. doi: 10.1093/aje/kws220
57. Gunn JS, Marshall JM, Baker S, Dongol S, Charles RC, Ryan ET. Salmonella chronic carriage: epidemiology, diagnosis, and gallbladder persistence. *Trends Microbiol.* (2014) 22:648–55. doi: 10.1016/j.tim.2014.06.007
58. Food E, Authority S. The European union summary report on trends and sources of zoonoses, zoonotic agents and food-borne outbreaks in 2017. *EFSA J.* (2018) 16:e05500. doi: 10.2903/j.efsa.2018.5500
59. Boscán-Duque L, Arzálluz-Fisher A, Ugarte C, Sánchez D, Wittum T, Hoet A. Reduced susceptibility to quinolones among salmonella serotypes isolated from poultry at slaughter in venezuela. *J Food Prot.* (2007) 70:2030–35. doi: 10.4315/0362-028X-70.9.2030
60. Medeiros MAN, de Oliveira DCN, Rodrigues DDP, de Freitas DRC. Prevalence and antimicrobial resistance of salmonella in chicken carcasses at retail in 15 Brazilian cities. *Rev Panam Salud Publica.* (2011) 30:555–60. doi: 10.1590/S1020-49892011001200010
61. Lapierre L, Cornejo J, Zavala S, Galarce N, Sánchez F, Benavides MB, et al. Phenotypic and genotypic characterization of virulence factors and susceptibility to antibiotics in salmonella infantis strains isolated from chicken meat: first findings in chile. *Animals.* (2020) 10:1049. doi: 10.3390/ani10061049
62. Foley SL, Nayak R, Hanning IB, Johnson TJ, Han J, Ricke SC. Population dynamics of salmonella enterica serotypes in commercial egg and poultry production. *Appl Environ Microbiol.* (2011) 77:4273–9. doi: 10.1128/AEM.00598-11
63. Da Cunha-Neto A, Carvalho LA, Carvalho RCT, Dos Prazeres Rodrigues D, Mano SB, De Souza Figueiredo EE, et al. Salmonella isolated from chicken carcasses from a slaughterhouse in the state of mato grosso, Brazil: antibiotic resistance profile, serotyping, and characterization by repetitive sequence-based PCR system. *Poult Sci.* (2018) 97:1373–81. doi: 10.3382/ps/pex406
64. Quesada A, Reginatto GA, Español AR, Colantonio LD, Burrone MS. Resistencia antimicrobiana de salmonella spp aislada de alimentos de origen animal para consumo humano. *Rev Peru Med Exp Salud Publica.* (2016) 33:32–44. doi: 10.17843/rpmesp.2016.331.1899
65. Rodriguez R, Fandiño C, Donado P, Guzmán L, Verjan N. Characterization of salmonella from commercial egg-laying hen farms in a central region of Colombia. *Avian Dis.* (2015) 59:57–63. doi: 10.1637/10873-052714-Reg
66. EFSA. The European union summary report on antimicrobial resistance in zoonotic and indicator bacteria from humans, animals and food in 2017/2018. *EFSA J.* (2020) 18:e06007. doi: 10.2903/j.efsa.2020.6007
67. Vinueza C. *Salmonella and Campylobacter in Broilers at Slaughter Age: A Possible Source for Carcasses Contamination in Ecuador*. Merelbeke: Faculty of Veterinary Medicine, Ghent University. (2017).
68. Van Boeckel TP, Brower C, Gilbert M, Grenfell BT, Levin S a, Robinson TP, et al. Global trends in antimicrobial use in food animals. *Proc Natl Acad Sci USA.* (2015) 112:5649–54. doi: 10.1073/pnas.1503141112
69. Hindermann D, Gopinath G, Chase H, Negrete F, Althaus D, Zurfluh K, et al. Salmonella enterica serovar infantis from food and human infections, Switzerland, 2010-2015: poultry-related multidrug resistant clones and an emerging ESBL producing clonal lineage. *Front Microbiol.* (2017) 8:1322. doi: 10.3389/fmicb.2017.01322

70. Bogomazova AN, Gordeeva VD, Krylova EV, Soltynskaya IV, Davydova EE, Ivanova OE, et al. Mega-plasmid found worldwide confers multiple antimicrobial resistance in salmonella infantis of broiler origin in Russia. *Int J Food Microbiol.* (2020) 319:108497. doi: 10.1016/j.ijfoodmicro.2019.108497
71. Wajid M, Saleemi MK, Sarwar Y, Ali A. Detection and characterization of multidrug-resistant salmonella enterica serovar infantis as an emerging threat in poultry farms of faisalabad, Pakistan. *J Appl Microbiol.* (2019) 127:248–61. doi: 10.1111/jam.14282
72. Vuthy Y, Lay KS, Seiha H, Kerleguer A, Aidara-Kane A. Antibiotic susceptibility and molecular characterization of resistance genes among *Escherichia coli* and among salmonella subsp. in chicken food chains. *Asian Pac J Trop Biomed.* (2017) 7:670–4. doi: 10.1016/j.apjtb.2017.07.002
73. Aviv G, Tsyba K, Steck N, Salmon-Divon M, Cornelius A, Rahav G, et al. A unique megaplasmid contributes to stress tolerance and pathogenicity of an emergent salmonella enterica serovar infantis strain. *Environ Microbiol.* (2014) 16:977–94. doi: 10.1111/1462-2920.12351
74. Franco A, Leekitcharoenphon P, Feltrin F, Alba P, Cordaro G, Iurescia M, et al. Emergence of a clonal lineage of multidrug-resistant ESBL-producing salmonella infantis transmitted from broilers and broiler meat to humans in Italy between 2011 and 2014. *PLoS ONE.* (2015) 10:e0144802. doi: 10.1371/journal.pone.0144802
75. Cohen E, Rahav G, Gal-Mor O. Genome sequence of an emerging salmonella enterica serovar infantis and genomic comparison with other S. infantis strains. *Genome Biol Evol.* (2020)12:223–8. doi: 10.1093/gbe/evaa048
76. Ledeboer NA, Frye JG, McClelland M, Jones BD. Salmonella enterica serovar typhimurium requires the Lpf, Pef, and Tafi fimbriae for biofilm formation on HEp-2 tissue culture cells and chicken intestinal epithelium. *Infect Immun.* (2006) 74:3156–69. doi: 10.1128/IAI.01428-05
77. Kingsley RA, Weening EH, Keestra AM, Bäumler AJ. Population heterogeneity of salmonella enterica serotype typhimurium resulting from phase variation of the lpf operon *in vitro* and *in vivo*. *J Bacteriol.* (2002) 184:2352–9. doi: 10.1128/JB.184.9.2352-2359.2002
78. Gupta SK, Sharma P, McMillan EA, Jackson CR, Hiott LM, Woodley T, et al. Genomic comparison of diverse salmonella serovars isolated from swine. *PLoS ONE.* (2019) 14:e0224518. doi: 10.1371/journal.pone.0224518
79. Boddicker JD, Ledeboer NA, Jagnow J, Jones BD, Clegg S. Differential binding to and biofilm formation on, HEp-2 cells by salmonella enterica serovar typhimurium is dependent upon allelic variation in the fimH gene of the fim gene cluster. *Mol Microbiol.* (2002) 45:1255–65. doi: 10.1046/j.1365-2958.2002.03121.x

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Mejía, Medina, Bayas, Salazar, Villavicencio, Zapata, Matheu, Wagenaar, González-Candelas and Vinueza-Burgos. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

REVIEW ARTICLE

GENOMIC MEDICINE

W. Gregory Feero, M.D., Ph.D., and Alan E. Guttmacher, M.D., *Editors*

Genomics, Type 2 Diabetes, and Obesity

Mark I. McCarthy, M.D.

From the Oxford Centre for Diabetes, Endocrinology and Metabolism; the Oxford National Institute of Health Research Biomedical Research Centre; and the Wellcome Trust Centre for Human Genetics, University of Oxford — all in Oxford, United Kingdom. Address reprint requests to Dr. McCarthy at the Oxford Centre for Diabetes, Endocrinology, and Metabolism, University of Oxford, Oxford OX3 7LJ, United Kingdom.

N Engl J Med 2010;363:2339-50.

Copyright © 2010 Massachusetts Medical Society.

TYPE 2 DIABETES, THOUGH POORLY UNDERSTOOD, IS KNOWN TO BE A DISease characterized by an inadequate beta-cell response to the progressive insulin resistance that typically accompanies advancing age, inactivity, and weight gain.[1] The disease accounts for substantial morbidity and mortality from adverse effects on cardiovascular risk and disease-specific complications such as blindness and renal failure.[2] The increasing global prevalence of type 2 diabetes is tied to rising rates of obesity[2] — in part a consequence of social trends toward higher energy intake and reduced energy expenditure. However, the mechanisms that underlie individual differences in the predisposition to obesity remain obscure.

Failure to understand the pathophysiology of diseases such as type 2 diabetes and obesity frustrates efforts to develop improved therapeutic and preventive strategies. The identification of DNA variants influencing disease predisposition will, it is hoped, deliver clues to the processes involved in disease pathogenesis. This would not only spur translational innovation but also provide opportunities for personalized medicine through stratification according to an individual person's risk and more precise classification of the disease subtype. In this article, I consider the extent to which these objectives have been realized.

DISCOVERY OF SUSCEPTIBILITY GENES

For type 2 diabetes and obesity, the discovery of causal genes (Fig. 1 and 2) has followed three main waves. The first wave consisted of family-based linkage analyses (see the Glossary) and focused candidate-gene studies. These proved effective in identifying genes responsible for extreme forms of early-onset disease segregating as single-gene (mendelian) disorders. Genes underlying several distinct, familial forms of nonautoimmune diabetes — including maturity-onset diabetes of the young, mitochondrial diabetes with deafness, and neonatal diabetes — were characterized (see the review by Waterfield and Gloyn[3]). Similar approaches revealed mutations in genes responsible for rare forms of severe childhood obesity, including the genes encoding leptin, the leptin receptor, and proopiomelanocortin (see the review by O'Rahilly[4]). These discoveries have provided insights into processes critical for the maintenance of normal glucose homeostasis and energy balance and clues to the inner workings of the pancreatic beta cell and hypothalamus. For many families, this information has led to improved diagnostic and therapeutic options (described in more detail below).

Attempts to apply similar approaches to families in which either common forms of diabetes or obesity is segregating have proved to be largely unrewarding,[5] and the second wave of discovery involved a switch to tests of association. Although intrinsically more powerful than linkage analysis, association analysis suffers from the disadvantage that the signal can be detected only if one examines

The New England Journal of Medicine
Downloaded from www.nejm.org at VIRGINIA COMMONWEALTH UNIV on December 9, 2010. For personal use only. No other uses without permission.
Copyright © 2010 Massachusetts Medical Society. All rights reserved.

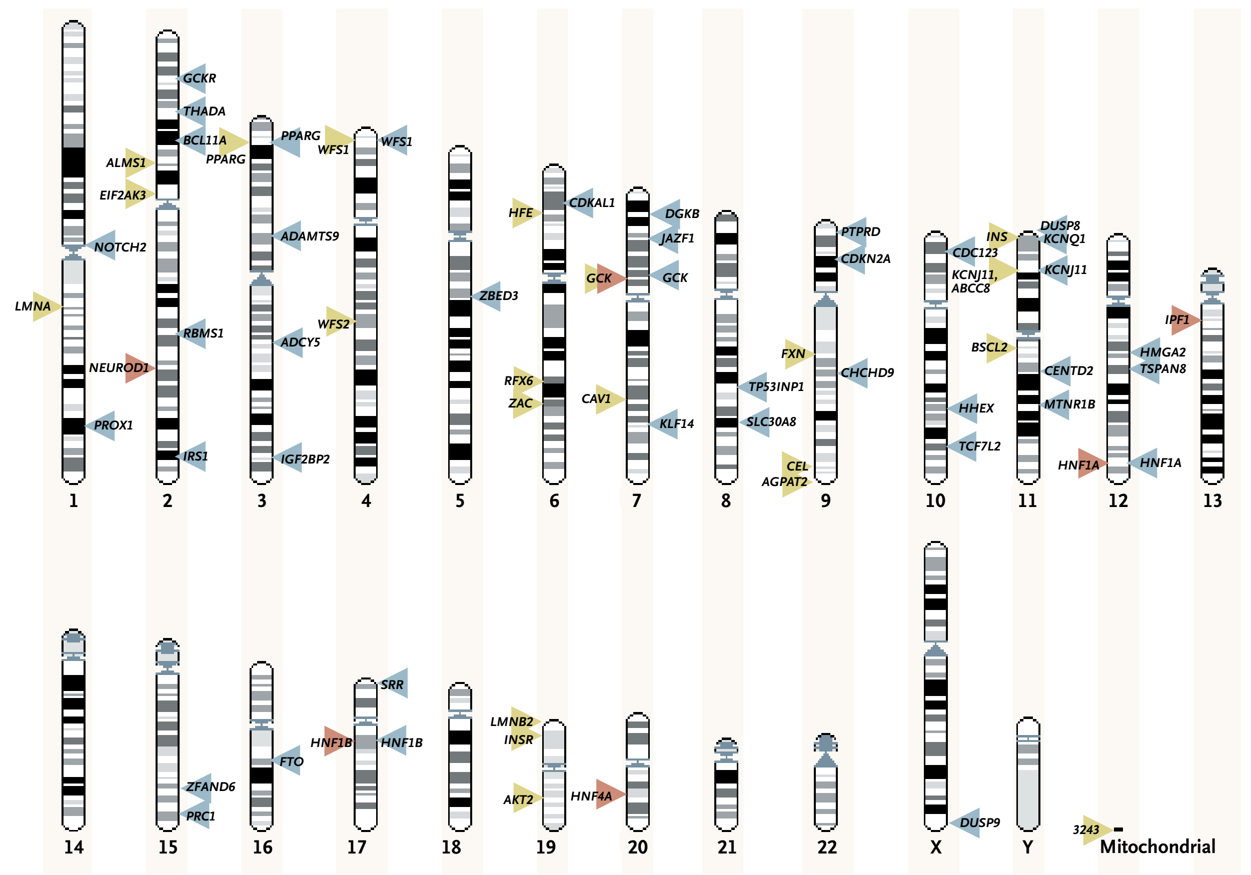

GCKR
THADA
BCL11A
PPARG
PPARG
WFS1
WFS1
ALMS1
EIF2AK3
NOTCH2
ADAMTS9
LMNA
WFS2
RBMS1
ADCY5
NEUROD1
PROX1
IRS1
IGF2BP2
ZBED3
HFE
CDKAL1
DGKB
JAZF1
GCK
GCK
RFX6
ZAC
CAV1
KLF14
TP53INP1
SLC30A8
FXN
CEL
AGPAT2
PTPRD
CDKN2A
CHCHD9
CDC123
KCNJ11,
ABCC8
INS
DUSP8
KCNQ1
KCNJ11
BSCL2
HHEX
TCF7L2
CENTD2
MTNR1B
HNF1A
HNF1A
HMGA2
TSPAN8
IPF1
1
2
3
4
5
6
7
8
9
10
11
12
13
ZFAND6
PRC1
FTO
HNF1B
HNF1B
SRR
LMNB2
INSR
AKT2
HNF4A
DUSP9
3243
Mitochondrial
14
15
16
17
18
19
20
21
22
X
Y

Figure 1. Genomic Locations of Proven Signals of Nonautoimmune Forms of Diabetes.

Signals are shown according to their location on each chromosome. Genes causing monogenic and selected syndromic forms of diabetes are shown to the left: genes implicated in maturity-onset diabetes of the young (red triangles) and those representing loci causal for other monogenic and syndromic forms of diabetes (green triangles). Common variants that have significant genomewide associations with multifactorial forms of diabetes are shown to the right (blue triangles); for these variants, the genes named within the triangles are indicative of signal position, but in most instances, formal proof that these are the specific genes responsible for the association is lacking.

the causal variant itself or a nearby marker with which it is tightly correlated. Until the advent of methods that enabled genomewide surveys of association, researchers were therefore obliged to direct their attention to specific candidate variants or genes of interest.[6] In retrospect, it is obvious that most such studies were seriously underpowered or focused on inappropriate candidates.[6] Nevertheless, by accruing data over the course of multiple studies, some genuine susceptibility variants were identified. Common coding variants in *PPARG* and *KCNJ11* (each of which encodes a protein that acts as a target for classes of therapeutic agents widely used in diabetes management) were shown to have modest effects on the risk of type 2 diabetes.[7,8] Resequencing of the gene encoding the melanocortin-4 receptor (MC4R) resulted in the identification of low-frequency coding variants that explain approximately 2 to 3% of cases of severe obesity.[9]

The third, and most successful, wave of discovery has been driven by systematic, large-scale surveys of association between common DNA sequence variants and disease. The first demonstration that unbiased discovery efforts could reveal new insights into the pathogenesis of type 2 diabetes resulted from identification of the association between type 2 diabetes and variants within *TCF7L2* (encoding transcription factor 7–like 2, a protein not previously identified as a biologic candidate).[10] *TCF7L2* has now been shown to modulate pancreatic islet function.[11]

The number of loci for which there is convincing evidence that they confer susceptibility to type 2 diabetes started to grow in early 2007 with the publication of the first genomewide as-

Downloaded from www.nejm.org at VIRGINIA COMMONWEALTH UNIV on December 9, 2010. For personal use only. No other uses without permission.
Copyright © 2010 Massachusetts Medical Society. All rights reserved.

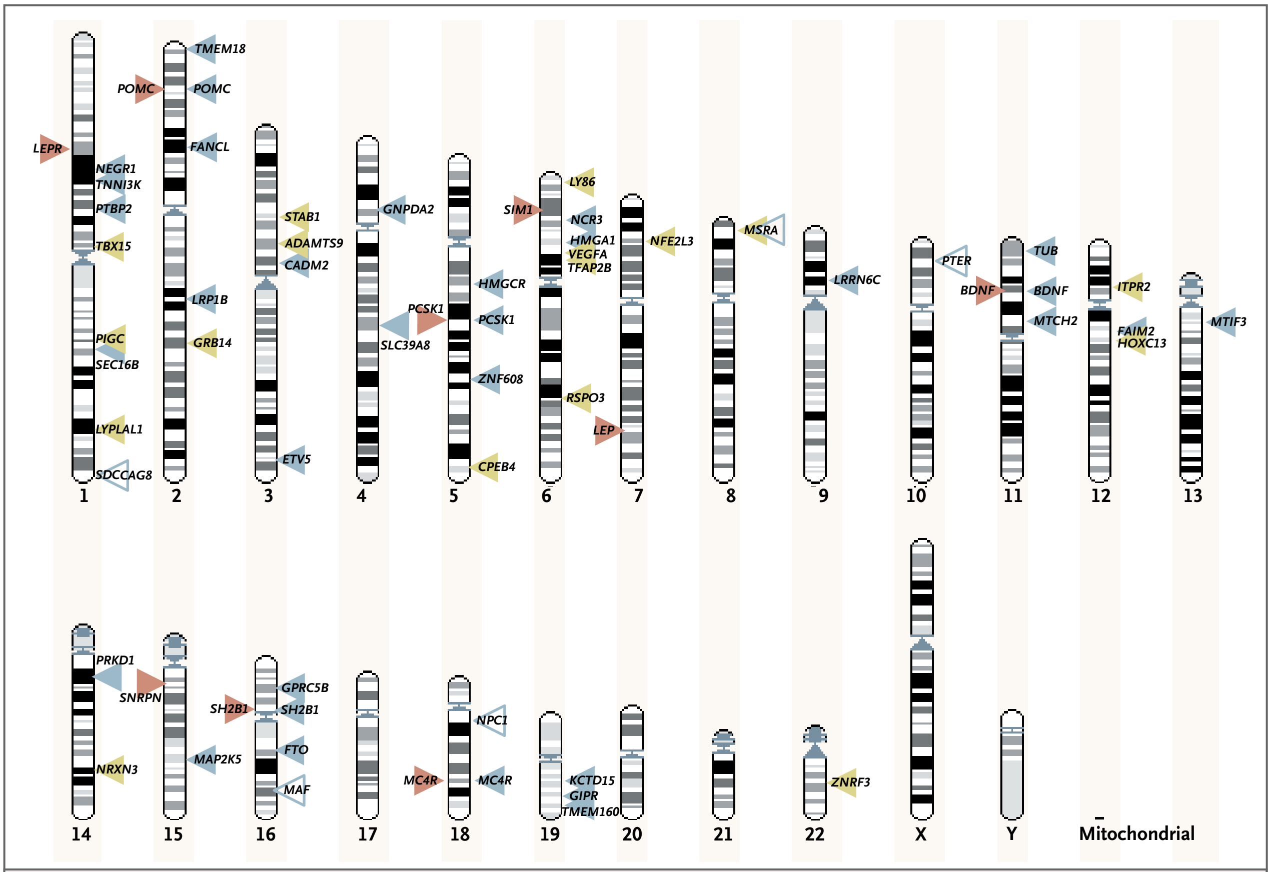

LEPR
POMC
TMEM18
POMC
FANCL
NEGR1
TNNI3K
PTBP2
TBX15
PIGC
SEC16B
LYPLAL1
SDCCAG8
LRP1B
GRB14
STAB1
ADAMTS9
CADM2
ETV5
GNPDA2
PCSK1
SLC39A8
HMGCR
PCSK1
ZNF608
CPEB4
SIM1
LY86
NCR3
HMGA1
VEGFA
TFAP2B
RSPO3
LEP
NFE2L3
MSRA
LRRN6C
PTER
BDNF
TUB
BDNF
MTCH2
ITPR2
FAIM2
HOXC13
MTIF3
1
2
3
4
5
6
7
8
9
10
11
12
13
PRKD1
SNRPN
NRXN3
MAP2K5
SH2B1
GPRC5B
SH2B1
FTO
MAF
NPC1
MC4R
MC4R
KCTD15
GIPR
TMEM160
ZNRF3
14
15
16
17
18
19
20
21
22
X
Y
Mitochondrial

Figure 2. Genomic Locations of Proven Signals of Body-Mass Index (BMI), Obesity, and Related Phenotypes.

Signals are shown according to their location on each chromosome. Genes causing monogenic and selected syndromic forms of obesity (red triangles) are shown to the left. Common variants that have significant genomewide associations with BMI or multifactorial obesity are shown to the right: loci implicated in BMI or weight variation at the population level (solid blue triangles), additional loci identified in case–control analyses of extreme obesity (open blue triangles), and variants identified primarily because of their association with waist circumference or waist-to-hip ratio (solid green triangles). For the variants shown to the right, the genes named within the triangles are indicative of signal position, but in most instances, formal proof that these are the specific genes responsible for the association is lacking.

sociation studies.[12-18] Together, these studies revealed six new associations, including variants near *CDKAL1*, *CDKN2A*, and *CDKN2B* (which encode putative or known regulators of cyclin-dependent kinases) and *HHEX* (which is transcribed into a homeobox protein implicated in beta-cell development). Typically each copy of a susceptibility allele at one of these loci is associated with a 15 to 20% increase in the risk of diabetes. Since then, the dominant approach to discovery has involved ever-larger aggregations of genomewide association data from multiple samples so as to improve the power to identify variants of modest effect: these studies have revealed more than 20 additional confirmed signals of susceptibility to type 2 diabetes[19-22] (Table 1 and Fig. 1). Though early studies were restricted to samples obtained from persons of European descent, genomewide association analyses conducted in other ethnic groups are now emerging.[23,24,29] The current total of approximately 40 confirmed type 2 diabetes loci includes variants in or near *WFS1* (wolframin) and the hepatocyte nuclear factors *HNF1A* and *HNF1B* (genes that also harbor rare mutations responsible for monogenic forms of diabetes)[30-33]; the melatonin-receptor gene *MTNR1B* (which highlights the link between circadian and metabolic regulation)[26-28]; and *IRS1* (encoding insulin-receptor substrate 1), one of a limited number of type 2 diabetes loci with a primary effect on insulin action rather than on secretion.[25]

Genomewide association studies of genetic variants influencing body-mass index (BMI) and

The New England Journal of Medicine
Downloaded from www.nejm.org at VIRGINIA COMMONWEALTH UNIV on December 9, 2010. For personal use only. No other uses without permission.
Copyright © 2010 Massachusetts Medical Society. All rights reserved.

Glossary

Allele: One of two or more versions of a genetic sequence at a particular location in the genome.

Association analysis: An approach to susceptibility-gene discovery that relies on identifying genetic variants whose allele frequencies are robustly correlated with either disease status or the level of a trait of interest.

Candidate-gene study: An approach to susceptibility-variant discovery that focuses on genetic analysis restricted to one or more candidate genes — genes that have typically been selected on the basis of a perceived match between their known or presumed functions and the biologic characteristics of the disease in question.

Coding variant: The part of the genomic DNA sequence that encodes proteins (consisting of approximately 1.5% of the total human genome).

Genomewide association study: An approach used in genetics research to look for associations between many (typically hundreds of thousands) of specific genetic variations (most commonly, single-nucleotide polymorphisms) and particular diseases or traits.

Homozygous: Having the same allele on both chromosomes at a particular location in the genome.

Linkage analysis: An approach to susceptibility-gene discovery that relies on matching family-level patterns of segregation of the disease of interest with genetic markers of known location.

Monogenic disease: Genetic disease attributable to variants with large effects on disease status. Because of the high penetrance of such variants, the disease typically cosegregates in a classic mendelian fashion (e.g., dominant or recessive).

Next-generation sequencing: DNA sequencing that harnesses advances in miniaturization technology to simultaneously sequence multiple areas of the genome rapidly and at low cost.

Noncoding variant: A DNA sequence variant that is located outside the coding sequence; some are likely to be involved in gene regulation.

Syndromic disease: Syndromes are characterized by the concomitant occurrence of several distinct clinical features. In syndromic forms of diabetes such as Wolfram's syndrome, a rare mutation of large effect leads not only to diabetes but also to a diversity of other features including optic atrophy and deafness.

Transcript: An RNA sequence resulting from transcription of a DNA sequence (often a gene).

obesity have been similarly productive, with three main strategies being adopted (Table 2 and Fig. 2). Genomewide association studies of population-based samples to examine the full range of BMI values have identified approximately 30 loci influencing BMI and the risk of obesity. The strongest signal remains the association with variants within *FTO* (the fat-mass and obesity–related gene).[13,34,45] Other signals near *BDNF*, *SH2B1*, and *NEGR1* (all implicated in aspects of neuronal function) reinforce the view of obesity as a disorder of hypothalamic function.[35,37,38,43] A second approach, focusing on case–control analysis of persons selected from the extremes of the BMI distribution, has delivered a complementary, only partly overlapping, set of loci.[39,42,46,47] Finally, genomewide analyses of patterns of fat distribution, prompted by the particularly deleterious health effects of visceral fat accumulation, have characterized approximately 15 loci that are largely distinct from those influencing overall adiposity[36,40,41,44]: many of the 15 display markedly stronger associations in women than in men.

FROM GENES TO CLINICAL PRACTICE

Despite the growing number of loci discovered, the contribution of genetic discoveries to the clinical management of diabetes and obesity remains limited to the small proportion of cases with monogenic forms of disease. What, then, are the obstacles impeding the clinical translation of the scores of multifactorial variants now defined?

The first is the modest effect size of the implicated variants. The common variants with the greatest effects on the risk of type 2 diabetes (*TCF7L2* in Europeans, *KCNQ1* in Asians) result in lifetime prevalence rates that are, in persons carrying two copies of the risk allele, roughly double those seen in persons with none.[10,23,24] The association signal at *FTO* accounts for less than 0.5% of the overall variance in BMI, equivalent to a difference of 2 to 3 kg between adults homozygous for the risk allele and those homozygous for the alternative allele.[13] Most other variants associated with type 2 diabetes and BMI have effects considerably smaller than these. More detailed analysis of the associated regions may reveal that some of these associations are driven by causal variants with larger effects, although empirical evidence supporting this assertion is limited.[22] In contrast, the mutations underlying monogenic forms of diabetes and obesity have far more dramatic clinical consequences: in pedigrees segregating these conditions, knowing whether a family member has inherited a given causal allele generally allows for the confident prediction of disease status.

A second obstacle to the translation of variants implicated in multifactorial forms of diabetes and obesity relates to the speed with which risk-allele discovery has led to an improved understanding of the biologic basis of disease. Most alleles implicated in monogenic and syndromic forms of diabetes and obesity alter the coding sequence and therefore have dramatic

The New England Journal of Medicine
Downloaded from www.nejm.org at VIRGINIA COMMONWEALTH UNIV on December 9, 2010. For personal use only. No other uses without permission.
Copyright © 2010 Massachusetts Medical Society. All rights reserved.

Table 1. Major Genomewide Association (GWA) Studies of Type 2 Diabetes.*

Reference	Sample Size		Major Ethnic Groups	Study Type	Main Findings
	GWA	Replication			
Sladek et al., 2007[12]	1,363	5,511	French	Single GWA study	*HHEX* and *SLC30A8* associations with type 2 diabetes
Scott et al., 2007[14]	2,335	2,473	Finnish	Single GWA study	*CDKAL1*, *CDKN2A*, and *IGF2BP2* associations with type 2 diabetes
Diabetes Genetics Initiative et al., 2007[15]	2,931	10,850	Swedish, Finnish	Single GWA study	*CDKAL1*, *CDKN2A*, and *IGF2BP2* associations with type 2 diabetes
Zeggini et al., 2007[16,18]	4,862	9,103	British	Single GWA study	*CDKAL1*, *CDKN2A*, and *IGF2BP2* associations with type 2 diabetes
Steinthorsdottir et al., 2007[17]	6,674	14,138	Icelandic	Single GWA study	*CDKAL1* association with type 2 diabetes and insulin secretion
Zeggini et al., 2008[19]	10,128	79,792	European	GWA meta-analysis	Six new loci for type 2 diabetes (*NOTCH2*, *JAZF1*, *ADAMTS9*, *TSPAN8*, *THADA*, and *CDC123*)
Yasuda et al., 2008[23]	1,691	18,239	Japanese, Korean, Chinese	Single GWA study	*KCNQ1* association with type 2 diabetes in East Asians
Unoki et al., 2008[24]	1,752	19,489	Japanese, Singaporean	Single GWA study	*KCNQ1* association with type 2 diabetes in East Asians
Rung et al., 2009[25]	1,376	27,033	French, Danish	Single GWA study	*IRS1* association with type 2 diabetes
Prokopenko et al., 2009[26]	36,610	82,689 For type 2 diabetes	European	Follow-up of signals for type 2 diabetes from GWA scan for fasting glucose	*MTNR1B* association with type 2 diabetes and fasting glucose
Lyssenko et al., 2009[27]	2,931	18,831 For type 2 diabetes	Swedish, Finnish	Follow-up of signals for type 2 diabetes from GWA scan for insulin secretion	*MTNR1B* association with type 2 diabetes and fasting glucose
Bouatia-Naji et al., 2009[28]	2,151	15,464 For type 2 diabetes	French, Danish, Finnish	Follow-up of signals for type 2 diabetes from GWA scan for fasting glucose	*MTNR1B* association with type 2 diabetes and fasting glucose
Dupuis et al., 2010[20]	46,186	127,677 For type 2 diabetes	European	Follow-up of signals for type 2 diabetes from GWA scan for fasting glucose	*ADCY5*, *PROX1*, *GCK*, *GCKR*, and *DGKB* associations with type 2 diabetes and fasting glucose
Tsai et al., 2010[29]	1,889	3,276	Taiwanese	Single GWA study	*SRR* and *PTPRD* associations with type 2 diabetes in Taiwanese
Qi et al., 2010[21]	5,643	84,605	European	GWA meta-analysis	*RBMS1* association with type 2 diabetes
Voight et al., 2010[22]	47,117	94,337	European	GWA meta-analysis	12 New loci for type 2 diabetes including *DUSP9*, *KLF14*, *CENTD2*, *HMGA2*, and *HNF1A*

* Only studies in which there were significant genomewide associations with type 2 diabetes are listed.

The New England Journal of Medicine
Downloaded from www.nejm.org at VIRGINIA COMMONWEALTH UNIV on December 9, 2010. For personal use only. No other uses without permission.
Copyright © 2010 Massachusetts Medical Society. All rights reserved.

Table 2. **Major Genomewide Association (GWA) Studies of Body-Mass Index (BMI), Risk of Obesity, and Fat Distribution.***

Reference	Sample Size		Major Ethnic Group	Study Type	Main Findings
	GWA	Replication			
Frayling et al., 2007[13]	4,862	38,759	British	GWA study of type 2 diabetes cohort	*FTO* association with BMI, obesity, and type 2 diabetes
Scuteri et al., 2007[34]	4,741	3,205	Sardinian	GWA study of large population isolate	*FTO* association with BMI
Loos et al., 2008[35]	16,876	75,981	European	GWA meta-analysis	Association of common *MC4R* variants with BMI
Chambers et al., 2008[36]	2,684	11,955	South Asian	GWA study of population sample	*MC4R* association with waist circumference
Willer et al., 2009[37]	32,387	59,082	European	GWA meta-analysis	*TMEM18, KCTD15, GNPDA2, SH2B1, MTCH2,* and *NEGR1* associations with BMI
Thorleifsson et al., 2009[38]	34,416	43,651	Icelandic	GWA study of large population isolate	Nine new associations with BMI (*NEGR1, TMEM18, ETV5, BDNF, FAIM2, KCTD15, SH2B1,* and *SEC16B*) or weight (*NCR3*)
Meyre et al., 2009[39]	2,796	14,186	French	GWA study of morbidly obese adults vs. normal-weight controls	Associations of signals near *NPC1, MAF,* and *PTER* with extreme obesity (from case–control analyses)
Lindgren et al., 2009[40]	38,580	70,639	European	GWA meta-analysis	Associations near *TFAP2B* and *MSRA* with waist circumference and near *LYPLAL1* with waist-to-hip ratio
Heard-Costa et al., 2009[41]	31,373	38,641	European	GWA meta-analysis	Associations of *NRXN3* variants with waist circumference
Scherag et al., 2010[42]	2,258	36,734	European	GWA study of persons with extreme early onset obesity vs. lean controls	*SDCCAG8* and *TNKS–MSRA* associations with childhood obesity
Speliotes et al., 2010[43]	123,865	125,931	European	GWA meta-analysis	18 New loci (including *POMC, GIPR,* and *HMGA1*) influencing BMI
Heid et al., 2010[44]	77,167	113,636	European	GWA meta-analysis	13 New loci (including *VEGFA, TBX15,* and *HOXC13*) influencing waist-to-hip ratio independent of BMI

* Only studies in which there were significant genomewide associations with BMI, waist circumference, waist-to-hip ratio, or obesity are listed.

The New England Journal of Medicine
Downloaded from www.nejm.org at VIRGINIA COMMONWEALTH UNIV on December 9, 2010. For personal use only. No other uses without permission.
Copyright © 2010 Massachusetts Medical Society. All rights reserved.

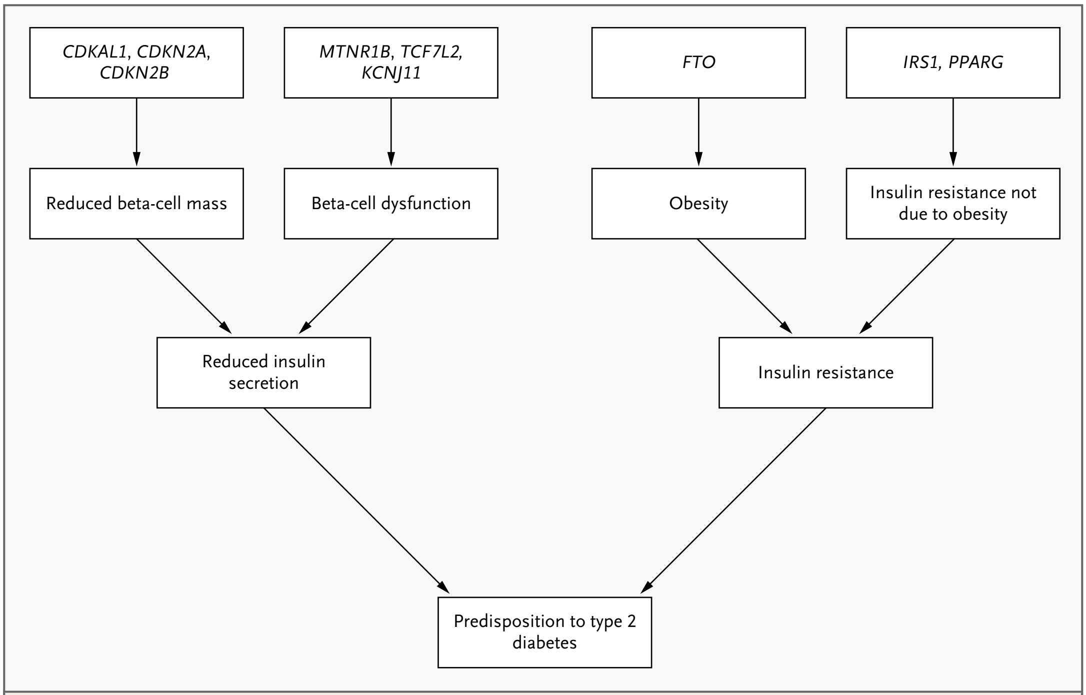


Figure 3. Pathways to Type 2 Diabetes Implicated by Identified Common Variant Associations.

Type 2 diabetes results when pancreatic beta cells are unable to secrete sufficient insulin to maintain normoglycemia, typically in the context of increasing peripheral insulin resistance. The beta-cell abnormalities fundamental to type 2 diabetes are thought to include both reduced beta-cell mass and disruptions of beta-cell function. Insulin resistance can be the consequence of obesity or of obesity-independent abnormalities in the responses of muscle, fat, or liver to insulin. Examples of susceptibility variants that, given current evidence, are likely to influence predisposition to type 2 diabetes by means of each of these mechanisms are shown.

and largely predictable effects on the function of the gene. The use of molecular diagnostics to derive clinically useful prognostic and therapeutic information relies on this relatively straightforward assignment of functional significance. In multifactorial disease, however, most susceptibility variants lie outside the coding regions of genes and are assumed to influence transcript regulation rather than gene function.

Characterization of the downstream consequences of these “noncoding” variants is difficult, given our rudimentary knowledge of the mechanics of gene regulation. Detailed functional studies are required to translate these genomic “signposts” into biologic knowledge that can spur translational development, and there have been relatively few successes.[48] Indeed, at most susceptibility loci, it remains far from clear even which transcripts mediate the susceptibility effects that have been observed.

The time required to achieve clinical translation is often underestimated,[49] and most of the discoveries in multifactorial disease have simply been too recent for their full translational potential to be realized. That potential lies in three main areas: the characterization of disease mechanisms that provide new targets for treatment and prevention, improved risk prediction and differential diagnosis, and personalized treatment and prevention.

FROM GENETICS TO BIOLOGY

An improved understanding of pathophysiology achieved through genetic discovery provides new opportunities for treatment, diagnosis, and monitoring. Studies of risk variants for type 2 diabetes in healthy populations have shown that most variants act through perturbation of insulin secretion rather than insulin action, establishing inherited abnormalities of beta-cell function or mass (or both) as critical components of the progression to type 2 diabetes (Fig. 3).[22,50] (An interactive graphic depicting proposed mechanisms of

An interactive graphic is available at NEJM.org

The New England Journal of Medicine
Downloaded from www.nejm.org at VIRGINIA COMMONWEALTH UNIV on December 9, 2010. For personal use only. No other uses without permission.
Copyright © 2010 Massachusetts Medical Society. All rights reserved.

some susceptibility variants associated with type 2 diabetes is available at NEJM.org.) At loci for which there is evidence of a primary effect driven by abnormalities of insulin action, both obesity-dependent and obesity-independent mechanisms are involved (Fig. 3).[22] As described above, it is not always easy to link association signals to specific transcripts, but some of the genes more confidently assigned to type 2 diabetes susceptibility — *TCF7L2*, *SLC30A8*, and *CDKN2A* and *CDKN2B* — relate to Wnt signaling, zinc transport, and cell-cycle regulation, respectively, suggesting that these functions have roles in the maintenance of normal islet function.[22,51] Beyond that, efforts to identify key processes in the pathogenesis of type 2 diabetes — for example, by showing that genes encoding members of particular pathways are overrepresented at susceptibility loci — have not been particularly rewarding.[22] Either type 2 diabetes is highly heterogeneous, or those fundamental disease processes are poorly captured by existing biologic knowledge.

Efforts to achieve therapeutic modification of weight have had little success. The identification of new pathways amenable to safe and effective weight manipulation would be a valuable "deliverable" from genetic-discovery efforts. However, the transition from association signal to causal mechanism has not been straightforward, especially when the disease involves tissues as inaccessible to direct study as the human hypothalamus. Consider the example of *FTO*.[13] Although the association signal maps to a clearly defined region of the gene, and the effect is comparatively large, there is still some doubt as to whether *FTO* itself is responsible for the weight phenotype, rather than one of the nearby genes such as *RPGRIP1L* (also expressed in the hypothalamus, with responses to alterations in nutritional and hormonal status similar to those of *FTO*[52]). Studies of mice with disruptions of *Fto* sequence[53,54] are consistent with the hypothesis that *FTO* mediates the BMI effect in humans, whereas studies of human *FTO* mutations have been less clear-cut.[55,56] Notwithstanding these data, the story emerging from the growing number of loci supports the view of overall obesity as a disease of hypothalamic dysregulation.[37,43] In contrast, variation in patterns of fat distribution is associated with variants within genes that influence adipocyte development and function.[40,41,44] How best to use this information to effect early translation into new therapeutic or preventive approaches remains uncertain.

One characteristic of metabolic disease is the cluster of traits referred to as the metabolic syndrome. However, the genetic evidence to date provides limited support for the metabolic syndrome as a defined pathophysiological entity, perhaps indicating that this clustering is driven by environmental factors. Though BMI-associated variants such as *FTO* modulate the risk of type 2 diabetes and hyperlipidemia,[57] and loci altering lipid levels have secondary effects on the risk of coronary artery disease,[58,59] there is little suggestion that the variants implicated in individual components of the metabolic syndrome overlap. At some loci, the patterns of association actually run counter to the broader correlative patterns of the metabolic syndrome. At the glucokinase regulator gene *GCKR*, for example, one common variant allele increases triglyceride levels yet lowers glucose levels.[15,60,61] The complexity of the relations that can exist at the genetic level between closely related phenotypes is further illustrated by the observation that alleles associated with similar degrees of fasting hyperglycemia in healthy populations have highly variable effects on the risk of type 2 diabetes later in life.[20]

PREDICTION AND DIFFERENTIAL DIAGNOSTICS

In cases of monogenic disease, genetic information can provide powerful diagnostic and predictive value for selected patients. Since subtypes of monogenic diabetes and obesity vary in their prognostic implications and therapeutic recommendations, a definitive molecular diagnosis is an important component of clinical management (Table 3).[3,62] To date, the use of molecular diagnostic tools has been limited by the expense of using conventional sequencing technologies to screen known causal genes for mutations that are often specific to a given family. Next-generation sequencing technologies are likely to be transformative in the medium term, though distinguishing pathogenic mutation from incidental variation will remain a challenge. In the meantime, improved biomarkers of diabetes subtypes that enable the more precise targeting of diagnostic resequencing would be valuable. For ex-

The New England Journal of Medicine
Downloaded from www.nejm.org at VIRGINIA COMMONWEALTH UNIV on December 9, 2010. For personal use only. No other uses without permission.
Copyright © 2010 Massachusetts Medical Society. All rights reserved.

Table 3. Initial Treatments for Various Diabetes Subtypes.*

Diabetes Subtype	Causal Genes	Optimal Treatment
Type 1 diabetes	About 40 known (genes in HLA region, *INS, PTPN22,* and others)	Lifelong insulin
Type 2 diabetes	About 40 known (*TCF7L2, CDKAL1,* and others)	Metformin as primary treatment; also sulfonylureas, glitazones, or insulin
LADA	Genes in HLA region, *INS,* and *PTPN22* (as in type 1 diabetes)	Early recourse to insulin therapy
GCK-MODY	*GCK*	Diet modification
HNF1A MODY	*HNF1A*	Sulfonylureas (low dose)
Mitochondrial diabetes	*MTTL1*	Early recourse to insulin therapy
Lipodystrophies	*LMNA, PPARG, AGPAT2, CAV1, BSCL2, LMNB2,* and *AKT2*	Uncertain; thiazolidinediones for some subtypes
Neonatal diabetes	*KCNJ11, ABCC8*	Sulfonylureas (high dose)
Neonatal diabetes	*INS*	Insulin

* GCK denotes glucokinase, LADA latent autoimmune diabetes in adults, MODY maturity-onset diabetes of the young, and tRNA transfer RNA.

ample, patients with maturity-onset diabetes of the young caused by *HNF1A* mutations have recently been shown to have C-reactive protein (CRP) levels well below those of patients with other subtypes of diabetes, suggesting that CRP could form the basis of a useful diagnostic test.[63] This observation also exemplifies the early translation of genetic discoveries, since it came directly from genomewide association studies showing that CRP levels are influenced by common variants near *HNF1A*.

The effect sizes of the known, common variants influencing the risk of type 2 diabetes and variation in adult BMI are modest, and the proportion of overall predisposition explained is small: approximately 5 to 10% for type 2 diabetes and 1% for BMI.[22,43] As a result, the ability to perform individual-level prediction with respect to these traits is limited. By combining data from multiple loci, one can identify persons who have inherited especially high or low numbers of risk alleles: the risk of type 2 diabetes differs by a factor of approximately 4 between persons in the top 1% and those in the bottom 1% of the "risk-score" distribution.[64-67] However, the risk profiles of many such persons are already discernible on the basis of conventional risk factors (e.g., BMI or family history), and there is limited evidence to suggest that information about genetic predisposition can be used effectively to guide the modification of long-term behavior. The discriminative accuracy of genetic profiling of known type 2 diabetes risk variants (as measured by means of receiver-operating-characteristic curves) is only approximately 60%,[64-67] well below the threshold required for clinical usefulness and the degree of prediction achievable on the basis of nongenetic risk factors.[68] Furthermore, estimates of risk can depend crucially on exactly which variants are included in the risk profile.[69] The key to improved performance will be the identification of risk variants with greater effect sizes than those discovered so far. Since existing genomewide association studies have most likely captured any common variants of large effect, the search is now focused on less-common variants.

A person's risk of type 2 diabetes or obesity reflects the joint effects of genetic predisposition and relevant environmental exposures. Efforts to determine whether these genetic and environmental components of risk interact (in the statistical sense that joint effects cannot be predicted from main effects alone)[70] face challenges associated with measuring relevant exposures (diet and physical activity being notoriously difficult to estimate) and the effect of imprecision on statistical power.[71] Although claims that statistical interactions reflect shared mechanisms (i.e., that the interacting factors act

The New England Journal of Medicine
Downloaded from www.nejm.org at VIRGINIA COMMONWEALTH UNIV on December 9, 2010. For personal use only. No other uses without permission.
Copyright © 2010 Massachusetts Medical Society. All rights reserved.

through the same pathways) are probably overstated, understanding the relative contributions of genetic and environmental components to risk is important. After all, environmental factors can be modified more readily than genetic factors.

Genetic discoveries have provided a molecular basis for the clinically useful classification of monogenic forms of diabetes and obesity.[3,4] Will the same be true for the common forms of these conditions? Probably not: as far as the common variants are concerned, each patient with diabetes or obesity has an individual "barcode" of susceptibility alleles and protective alleles across many loci. It is possible to show that the genetic profiles of lean subjects with type 2 diabetes and obese subjects with type 2 diabetes are not identical, but these differences appear to be inadequate for clinically useful subclassification.[22,72] If efforts to uncover less prevalent, higher-penetrance alleles are successful, more precise classification of disease subtypes may become possible, particularly if genetic data can be integrated with clinical and biochemical information. For example, in persons presenting with diabetes in early adulthood, there are several possible diagnoses: various subtypes of maturity-onset diabetes of the young or mitochondrial diabetes, for example, as well as type 1 or type 2 diabetes. Assigning the correct diagnosis has both prognostic and therapeutic benefits for the patient (Table 3).

TARGETED TREATMENT AND PREVENTION

Recommended therapies for the various subtypes of diabetes differ (Table 3).[3,4,62,73-75] In monogenic forms of diabetes, at least, genetic testing already drives the choice of therapy. For example, in patients who have maturity-onset diabetes of the young due to mutations in the gene encoding glucokinase (GCK), the hyperglycemia is mild and stable, the risk of complications is low, and dietary management is often sufficient. In contrast, in patients who have maturity-onset diabetes of the young due to mutations in *HNF1A*, the disease follows a more aggressive course, with a greater risk of severe complications, but is particularly responsive to the hypoglycemic effects of sulfonylureas.[62,73] Most children with neonatal diabetes have mutations in *KCNJ11* or *ABCC8*, adjacent genes that jointly encode the beta-cell ATP-sensitive potassium channel that mediates glucose-stimulated insulin secretion and is the target of sulfonylureas. In such children, treatment with sulfonylureas has proved more effective and convenient than the lifelong insulin therapy previously considered the default option.[74,75] In children with severe obesity due to profound leptin deficiency, exogenous leptin therapy is lifesaving.[76]

As yet, there are insufficient genetic data to support management decisions for common forms of type 2 diabetes and obesity.[77] Although the *TCF7L2* genotype is associated with variation in the response to sulfonylurea treatment,[78] the effect is too modest to guide the care of individual patients. For the time being, the contribution of genetic information to therapy is most likely to come through the drug-discovery pipeline. Information from genetic studies could be used to identify new targets for pharmaceutical intervention that have validated effects on physiological characteristics, to provide information about new and existing targets (e.g., clues about the long-term safety of pathway intervention),[32] and to characterize high-risk groups to enable more efficient clinical trials of agents designed to reduce the progression of type 2 diabetes or obesity or the risk of complications.

SUMMARY

Given the substantial time it takes to translate basic biomedical discoveries into clinical tools,[49] any current assessment of the clinical value of recent advances in the genetic basis of common diseases is probably an underestimate. An improved understanding of fundamental disease mechanisms is already emerging; this will underpin future therapeutic advances. But the expansion of personalized medicine beyond monogenic forms of disease awaits a more complete description of predisposition. The boundaries of personalized medicine will be much clearer in a few years, after large-scale genomewide resequencing efforts (now under way) provide a systematic, comprehensive description of the relations between genome sequence variation and major clinical phenotypes.

Dr. McCarthy reports receiving consulting fees from Prosidion Pharmaceuticals and lecture fees from Novo Nordisk. No other potential conflict of interest relevant to this article was reported.

Disclosure forms provided by the author are available with the full text of this article at NEJM.org.

The New England Journal of Medicine
Downloaded from www.nejm.org at VIRGINIA COMMONWEALTH UNIV on December 9, 2010. For personal use only. No other uses without permission.
Copyright © 2010 Massachusetts Medical Society. All rights reserved.

REFERENCES

1. Stumvoll M, Goldstein BJ, van Haeften TW. Type 2 diabetes: pathogenesis and treatment. Lancet 2008;371:2153-6.

2. Zimmet P, Alberti KG, Shaw J. Global and societal implications of the diabetes epidemic. Nature 2001;414:782-7.

3. Waterfield T, Gloyn AL. Monogenic β-cell dysfunction in children: clinical phenotypes, genetic etiology and mutational pathways. Pediatr Health 2008;2:517-32.

4. O'Rahilly S. Human genetics illuminates the paths to metabolic disease. Nature 2009;462:307-14.

5. McCarthy MI. Growing evidence for diabetes susceptibility genes from genome scan data. Curr Diab Rep 2003;3:159-67.

6. Hattersley AT, McCarthy MI. What makes a good genetic association study? Lancet 2005;366:1315-23.

7. Altshuler D, Hirschhorn JN, Klannemark M, et al. The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. Nat Genet 2000;26:76-80.

8. Gloyn AL, Weedon MN, Owen KR, et al. Large-scale association studies of variants in genes encoding the pancreatic beta-cell KATP channel subunits Kir6.2 (KCNJ11) and SUR1 (ABCC8) confirm that the KCNJ11 E23K variant is associated with type 2 diabetes. Diabetes 2003;52:568-72.

9. Larsen LH, Echwald SM, Sørensen TI, Andersen T, Wulff BS, Pedersen O. Prevalence of mutations and functional analyses of melanocortin 4 receptor variants identified among 750 men with juvenile-onset obesity. J Clin Endocrinol Metab 2005;90:219-24.

10. Grant SF, Thorleifsson G, Reynisdottir I, et al. Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. Nat Genet 2006;38:320-3.

11. Lyssenko V, Lupi R, Marchetti P, et al. Mechanisms by which common variants in the TCF7L2 gene increase risk of type 2 diabetes. J Clin Invest 2007;117:2155-63.

12. Sladek R, Rocheleau G, Rung J, et al. A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature 2007;445:881-5.

13. Frayling TM, Timpson NJ, Weedon MN, et al. A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. Science 2007;316:889-94.

14. Scott LJ, Mohlke KL, Bonnycastle LL, et al. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. Science 2007;316:1341-5.

15. Diabetes Genetics Initiative of Broad Institute of Harvard and MIT, Lund University, Novartis Institutes of BioMedical Research, et al. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. Science 2007;316:1331-6.

16. Zeggini E, Weedon MN, Lindgren CM, et al. Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. Science 2007;316:1336-41. [Erratum, Science 2007;317:1035-6.]

17. Steinthorsdottir V, Thorleifsson G, Reynisdottir I, et al. A variant in CDKAL1 influences insulin response and risk of type 2 diabetes. Nat Genet 2007;39:770-5.

18. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 2007;447:661-78.

19. Zeggini E, Scott LJ, Saxena R, et al. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. Nat Genet 2008;40:638-45.

20. Dupuis J, Langenberg C, Prokopenko I, et al. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. Nat Genet 2010;42:105-16.

21. Qi L, Cornelis MC, Kraft P, et al. Genetic variants at 2q24 are associated with susceptibility to type 2 diabetes. Hum Mol Genet 2010;19:2706-15.

22. Voight BF, Scott LJ, Steinthorsdottir V, et al. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. Nat Genet 2010;42:579-89.

23. Yasuda K, Miyake K, Horikawa Y, et al. Variants in KCNQ1 are associated with susceptibility to type 2 diabetes mellitus. Nat Genet 2008;40:1092-7.

24. Unoki H, Takahashi A, Kawaguchi T, et al. SNPs in KCNQ1 are associated with susceptibility to type 2 diabetes in East Asian and European populations. Nat Genet 2008;40:1098-102.

25. Rung J, Cauchi S, Albrechtsen A, et al. Genetic variant near IRS1 is associated with type 2 diabetes, insulin resistance and hyperinsulinemia. Nat Genet 2009;41:1110-5. [Erratum, Nat Genet 2009;41:1156.]

26. Prokopenko I, Langenberg C, Florez JC, et al. Variants in MTNR1B influence fasting glucose levels. Nat Genet 2009;41:77-81.

27. Lyssenko V, Nagorny CL, Erdos MR, et al. Common variant in MTNR1B associated with increased risk of type 2 diabetes and impaired early insulin secretion. Nat Genet 2009;41:82-8.

28. Bouatia-Naji N, Bonnefond A, Cavalcanti-Proença C, et al. A variant near MTNR1B is associated with increased fasting plasma glucose levels and type 2 diabetes risk. Nat Genet 2009;41:89-94.

29. Tsai FJ, Yang CF, Chen CC, et al. A genome-wide association study identifies susceptibility variants for type 2 diabetes in Han Chinese. PLoS Genet 2010;6(2):e1000847.

30. Sandhu MS, Weedon MN, Fawcett KA, et al. Common variants in WFS1 confer risk of type 2 diabetes. Nat Genet 2007;39:951-3.

31. Winckler W, Weedon MN, Graham RR, et al. Evaluation of common variants in the six known maturity-onset diabetes of the young (MODY) genes for association with type 2 diabetes. Diabetes 2007;56:685-93.

32. Gudmundsson J, Sulem P, Steinthorsdottir V, et al. Two variants on chromosome 17 confer prostate cancer risk, and the one in TCF2 protects against type 2 diabetes. Nat Genet 2007;39:977-83.

33. Franks PW, Rolandsson O, Debenham SL, et al. Replication of the association between variants in WFS1 and risk of type 2 diabetes in European populations. Diabetologia 2008;51:458-63. [Erratum, Diabetologia 2008;51:523.]

34. Scuteri A, Sanna S, Chen WM, et al. Genome-wide association scan shows genetic variants in the FTO gene are associated with obesity-related traits. PLoS Genet 2007;3(7):e115.

35. Loos RJ, Lindgren CM, Li S, et al. Common variants near MC4R are associated with fat mass, weight and risk of obesity. Nat Genet 2008;40:768-75.

36. Chambers JC, Elliott P, Zabaneh D, et al. Common genetic variation near MC4R is associated with waist circumference and insulin resistance. Nat Genet 2008;40:716-8.

37. Willer CJ, Speliotes EK, Loos RJ, et al. Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. Nat Genet 2009;41:25-34.

38. Thorleifsson G, Walters GB, Gudbjartsson DF, et al. Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity. Nat Genet 2009;41:18-24.

39. Meyre D, Delplanque J, Chèvre JC, et al. Genome-wide association study for early-onset and morbid adult obesity identifies three new risk loci in European populations. Nat Genet 2009;41:157-9.

40. Lindgren CM, Heid IM, Randall JC, et al. Genome-wide association scan meta-analysis identifies three loci influencing adiposity and fat distribution. PLoS Genet 2009;5(6):e1000508.

41. Heard-Costa NL, Zillikens MC, Monda KL, et al. NRXN3 is a novel locus for waist circumference: a genome-wide association study from the CHARGE Consortium. PLoS Genet 2009;5(6):e1000539.

42. Scherag A, Dina C, Hinney A, et al.

The New England Journal of Medicine
Downloaded from www.nejm.org at VIRGINIA COMMONWEALTH UNIV on December 9, 2010. For personal use only. No other uses without permission.
Copyright © 2010 Massachusetts Medical Society. All rights reserved.

Two new loci for body-weight regulation identified in a joint analysis of genome-wide association studies for early-onset extreme obesity in French and German study groups. PLoS Genet 2010;6(4):e1000916.
43. Speliotes EK, Willer CJ, Berndt SI, et al. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nat Genet 2010;42:937-48.
44. Heid IM, Jackson AU, Randall JC, et al. Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution. Nat Genet 2010;42:949-60.
45. Dina C, Meyre D, Gallina S, et al. Variation in FTO contributes to childhood obesity and severe adult obesity. Nat Genet 2007;39:724-6.
46. Benzinou M, Creemers JW, Choquet H, et al. Common nonsynonymous variants in PCSK1 confer risk of obesity. Nat Genet 2008;40:943-5.
47. Walters RG, Jacquemont S, Valsesia A, et al. A new highly penetrant form of obesity due to deletions on chromosome 16p11.2. Nature 2010;463:671-5.
48. Gaulton KJ, Nammo T, Pasquali L, et al. A map of open chromatin in human pancreatic islets. Nat Genet 2010;42:255-9.
49. Contopoulos-Ioannidis DG, Alexiou GA, Gouvias TC, Ioannidis JP. Life cycle of translational research for medical interventions. Science 2008;321:1298-9.
50. Florez JC. Newly identified loci highlight beta cell dysfunction as a key cause of type 2 diabetes: where are the insulin resistance genes? Diabetologia 2008;51:1100-10.
51. Prokopenko I, McCarthy MI, Lindgren CM. Type 2 diabetes: new genes, new understanding. Trends Genet 2008;24:613-21.
52. Stratigopoulos G, Padilla SL, LeDuc CA, et al. Regulation of Fto/Ftm gene expression in mice and humans. Am J Physiol Regul Integr Comp Physiol 2008;294:R1185-R1196. [Erratum, Am J Physiol Regul Integr Comp Physiol 2008;295:R1360-R1363.]
53. Fischer J, Koch L, Emmerling C, et al. Inactivation of the Fto gene protects from obesity. Nature 2009;458:894-8.
54. Church C, Lee S, Bagg EA, et al. A mouse model for the metabolic effects of the human fat mass and obesity associated FTO gene. PLoS Genet 2009;5(8):e1000599.
55. Boissel S, Reish O, Proulx K, et al. Loss-of-function mutation in the dioxygenase-encoding FTO gene causes severe growth retardation and multiple malformations. Am J Hum Genet 2009;85:106-11.
56. Meyre D, Proulx K, Kawagoe-Takaki H, et al. Prevalence of loss-of-function FTO mutations in lean and obese individuals. Diabetes 2010;59:311-8.
57. Freathy RM, Timpson NJ, Lawlor DA, et al. Common variation in the FTO gene alters diabetes-related metabolic traits to the extent expected given its effect on BMI. Diabetes 2008;57:1419-26.
58. Kathiresan S, Willer CJ, Peloso GM, et al. Common variants at 30 loci contribute to polygenic dyslipidemia. Nat Genet 2009;41:56-65.
59. Myocardial Infarction Genetics Consortium, Kathiresan S, Voight BF, et al. Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. Nat Genet 2009;41:334-41.
60. Orho-Melander M, Melander O, Guiducci C, et al. Common missense variant in the glucokinase regulatory protein gene is associated with increased plasma triglyceride and C-reactive protein but lower fasting glucose concentrations. Diabetes 2008;57:3112-21.
61. Vaxillaire M, Cavalcanti-Proença C, Dechaume A, et al. The common P446L polymorphism in GCKR inversely modulates fasting glucose and triglyceride levels and reduces type 2 diabetes risk in the DESIR prospective general French population. Diabetes 2008;57:2253-7.
62. Murphy R, Ellard S, Hattersley AT. Clinical implications of a molecular genetic classification of monogenic beta-cell diabetes. Nat Clin Pract Endocrinol Metab 2008;4:200-13.
63. Owen KR, Thanabalasingham G, James TJ, et al. Assessment of highly sensitive C-reactive protein levels as diagnostic discriminator of maturity onset diabetes of the young due to HNF1A mutations. Diabetes Care 2010;33:1919-24.
64. Lango H, UK Type 2 Diabetes Genetics Consortium, Palmer CN, et al. Assessing the combined impact of 18 common genetic variants of modest effect sizes on type 2 diabetes risk. Diabetes 2008;57:3129-35.
65. Lyssenko V, Jonsson A, Almgren P, et al. Clinical risk factors, DNA variants, and the development of type 2 diabetes. N Engl J Med 2008;359:2220-32.
66. Cornelis MC, Qi L, Zhang C, et al. Joint effects of common genetic variants on the risk for type 2 diabetes in U.S. men and women of European ancestry. Ann Intern Med 2009;150:541-50.
67. van Hoek M, Dehghan A, Witteman JC, et al. Predicting type 2 diabetes based on polymorphisms from genome-wide association studies: a population-based study. Diabetes 2008;57:3122-8.
68. Janssens AC, van Duijn CM. Genome-based prediction of common diseases: advances and prospects. Hum Mol Genet 2008;17:R166-R173.
69. Mihaescu R, van Hoek M, Sijbrands EJ, et al. Evaluation of risk prediction updates from commercial genome-wide scans. Genet Med 2009;11:588-94.
70. Wareham NJ, Franks PW, Harding AH. Establishing the role of gene-environment interactions in the etiology of type 2 diabetes. Endocrinol Metab Clin North Am 2002;31:553-66.
71. Luan JA, Wong MY, Day NE, Wareham NJ. Sample size determination for studies of gene-environment interaction. Int J Epidemiol 2001;30:1035-40.
72. Timpson NJ, Lindgren CM, Weedon MN, et al. Adiposity-related heterogeneity in patterns of type 2 diabetes susceptibility observed in genome-wide association data. Diabetes 2009;58:505-10.
73. Pearson ER, Starkey BJ, Powell RJ, Gribble FM, Clark PM, Hattersley AT. Genetic cause of hyperglycaemia and response to treatment in diabetes. Lancet 2003;362:1275-81.
74. Pearson ER, Flechtner I, Njølstad PR, et al. Switching from insulin to oral sulfonylureas in patients with diabetes due to Kir6.2 mutations. N Engl J Med 2006;355:467-77.
75. Rafiq M, Flanagan SE, Patch AM, et al. Effective treatment with oral sulfonylureas in patients with diabetes due to sulfonylurea receptor 1 (SUR1) mutations. Diabetes Care 2008;31:204-9.
76. Farooqi IS, Jebb SA, Langmack G, et al. Effects of recombinant leptin therapy in a child with congenital leptin deficiency. N Engl J Med 1999;341:879-84.
77. Pearson ER. Pharmacogenetics in diabetes. Curr Diab Rep 2009;9:172-81.
78. Pearson ER, Donnelly LA, Kimber C, et al. Variation in TCF7L2 influences therapeutic response to sulfonylureas: a GoDARTs study. Diabetes 2007;56:2178-82.

Copyright © 2010 Massachusetts Medical Society.

The New England Journal of Medicine
Downloaded from www.nejm.org at VIRGINIA COMMONWEALTH UNIV on December 9, 2010. For personal use only. No other uses without permission.
Copyright © 2010 Massachusetts Medical Society. All rights reserved.

REVIEW ARTICLE

GENOMIC MEDICINE

W. Gregory Feero, M.D., Ph.D., and Alan E. Guttmacher, M.D., *Editors*

Genomics of Cardiovascular Disease

Christopher J. O'Donnell, M.D., and Elizabeth G. Nabel, M.D.

From the National Heart, Lung, and Blood Institute and the Framingham Heart Study, National Institutes of Health, Bethesda, MD (C.J.O.); and Massachusetts General Hospital (C.J.O.), Brigham and Women's Hospital (E.G.N.), and Harvard Medical School (C.J.O., E.G.N.) — all in Boston. Address reprint requests to Dr. Nabel at Brigham and Women's Hospital, 75 Francis St., Boston, MA 02115, or at enabel@partners.org.

N Engl J Med 2011;365:2098-109.

Copyright © 2011 Massachusetts Medical Society.

An interactive timeline is available at NEJM.org

CARDIOVASCULAR DISEASE IS THE LEADING CAUSE OF DEATH IN THE UNITED States. Considerable progress has been made in the past 50 years to define, identify, and modify risk factors for cardiovascular disease (e.g., hypertension, dyslipidemia, obesity, type 2 diabetes, cigarette smoking, and physical inactivity) and to develop treatments, such as coronary care units, percutaneous coronary interventions, and beta-blockers. These efforts have resulted in an age-adjusted decline in cardiovascular mortality.[1] In addition, it is now possible to detect subclinical disease by means of blood biomarker testing[2] or imaging measurements[3] years before the onset of symptoms or other clinical manifestations.[4] Despite this progress, mechanisms that underlie individual differences in the presentation and pathophysiological features of cardiovascular disease are poorly understood. In this article, we review genetic and genomic studies in cardiovascular medicine that have helped to elucidate some of these mechanisms during the past decade (Fig. 1, and interactive timeline, available with the full text of this article at NEJM.org).

MENDELIAN AND CANDIDATE-GENE STUDIES

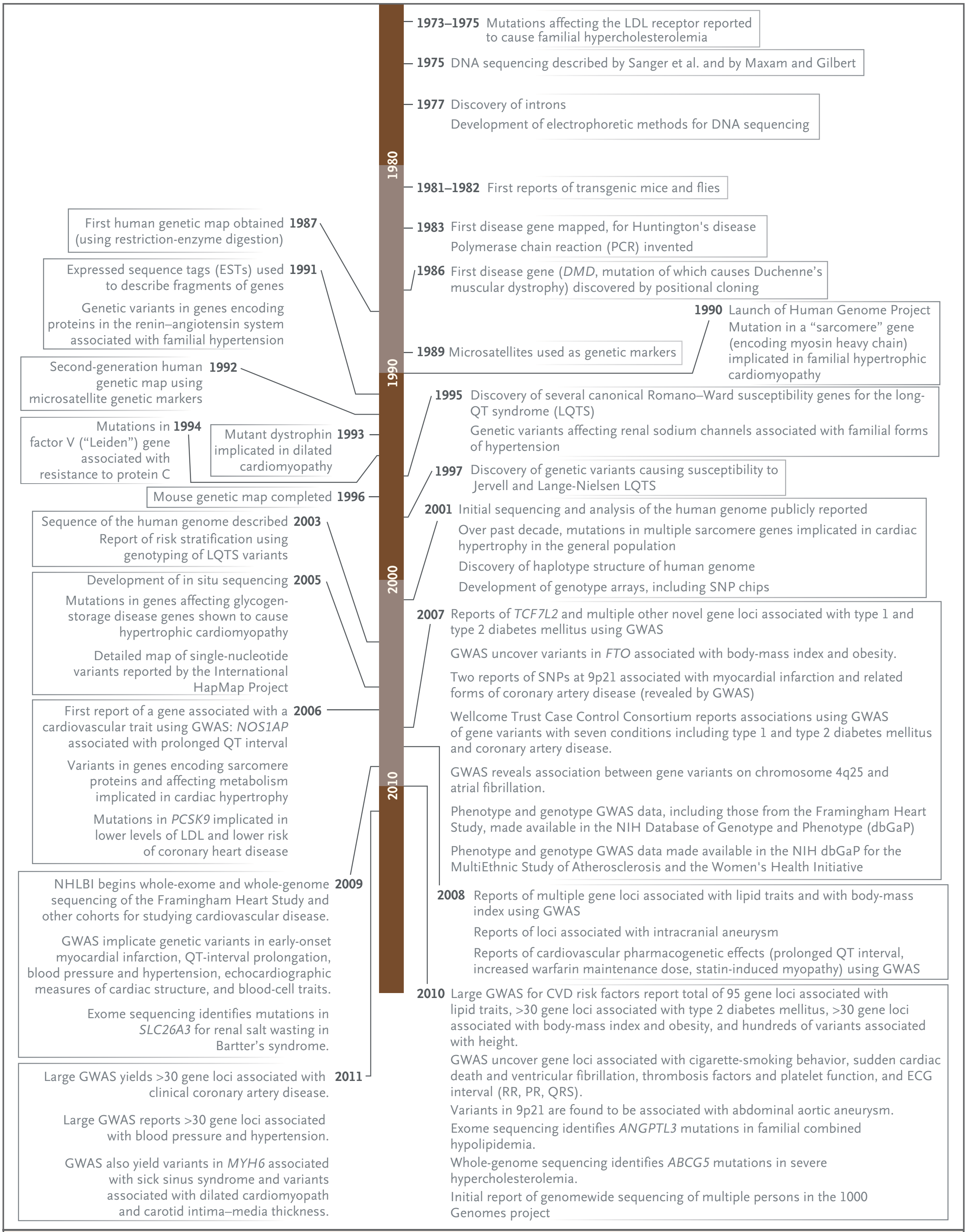
Ten years ago, the draft sequence of the human genome, which was produced by scientists working on the Human Genome Project[5] and others,[6] was first described, leading to an expansion in the understanding of genetic contributions to cardiovascular disease. Before the Human Genome Project, many genes associated with mendelian cardiovascular disease had been identified.[7] These forms of cardiovascular disease are rare and constitute a minority of clinical cardiovascular diseases. Genetically, they are simple in that a mutation in a single gene is sufficient to cause disease, so mendelian disease is said to be monogenic. Examples include forms of premature myocardial infarction, dilated and hypertrophic cardiomyopathy, heart failure, arrhythmogenic right ventricular dysplasia, the long-QT syndrome, and aortic aneurysms.[7] Recessive mutations underlie familial forms of cardiovascular risk factors, such as hypertension, hypercholesterolemia, and type 2 diabetes. Knowledge obtained through the identification of genes associated with mendelian disease has led to breakthrough discoveries in mechanisms of cardiovascular disease and its treatment. A compelling illustration is the Nobel-Prize–winning discovery that mutations affecting the low-density lipoprotein (LDL) receptor cause hypercholesterolemia and early-onset myocardial infarction, which led to LDL cholesterol–lowering therapies that reduce the risk of cardiovascular events.[8]

GENOMICS OF CARDIOVASCULAR RISK AND DISEASE

ROLE OF GENOMEWIDE ASSOCIATION STUDIES

The large majority of cardiovascular diseases, however, are polygenic, with both heritable and environmental contributions.[9] Moreover, there are heritable, polygenic com-

The New England Journal of Medicine
Downloaded from nejm.org at VIRGINIA COMMONWEALTH UNIV on December 1, 2011. For personal use only. No other uses without permission.
Copyright © 2011 Massachusetts Medical Society. All rights reserved.



1973–1975 Mutations affecting the LDL receptor reported to cause familial hypercholesterolemia

1975 DNA sequencing described by Sanger et al. and by Maxam and Gilbert

1977 Discovery of introns

Development of electrophoretic methods for DNA sequencing

1981–1982 First reports of transgenic mice and flies

1983 First disease gene mapped, for Huntington's disease

Polymerase chain reaction (PCR) invented

1986 First disease gene (*DMD*, mutation of which causes Duchenne's muscular dystrophy) discovered by positional cloning

First human genetic map obtained (using restriction-enzyme digestion) **1987**

1989 Microsatellites used as genetic markers

1990 Launch of Human Genome Project

Mutation in a "sarcomere" gene (encoding myosin heavy chain) implicated in familial hypertrophic cardiomyopathy

Expressed sequence tags (ESTs) used to describe fragments of genes **1991**

Genetic variants in genes encoding proteins in the renin–angiotensin system associated with familial hypertension

Second-generation human genetic map using microsatellite genetic markers **1992**

Mutant dystrophin implicated in dilated cardiomyopathy **1993**

Mutations in factor V ("Leiden") gene associated with resistance to protein C **1994**

1995 Discovery of several canonical Romano–Ward susceptibility genes for the long-QT syndrome (LQTS)

Genetic variants affecting renal sodium channels associated with familial forms of hypertension

Mouse genetic map completed **1996**

1997 Discovery of genetic variants causing susceptibility to Jervell and Lange-Nielsen LQTS

2001 Initial sequencing and analysis of the human genome publicly reported

Over past decade, mutations in multiple sarcomere genes implicated in cardiac hypertrophy in the general population

Discovery of haplotype structure of human genome

Development of genotype arrays, including SNP chips

Sequence of the human genome described **2003**

Report of risk stratification using genotyping of LQTS variants

Development of in situ sequencing **2005**

Mutations in genes affecting glycogen-storage disease genes shown to cause hypertrophic cardiomyopathy

Detailed map of single-nucleotide variants reported by the International HapMap Project

First report of a gene associated with a cardiovascular trait using GWAS: *NOS1AP* associated with prolonged QT interval **2006**

Variants in genes encoding sarcomere proteins and affecting metabolism implicated in cardiac hypertrophy

Mutations in *PCSK9* implicated in lower levels of LDL and lower risk of coronary heart disease

2007 Reports of *TCF7L2* and multiple other novel gene loci associated with type 1 and type 2 diabetes mellitus using GWAS

GWAS uncover variants in *FTO* associated with body-mass index and obesity.

Two reports of SNPs at 9p21 associated with myocardial infarction and related forms of coronary artery disease (revealed by GWAS)

Wellcome Trust Case Control Consortium reports associations using GWAS of gene variants with seven conditions including type 1 and type 2 diabetes mellitus and coronary artery disease.

GWAS reveals association between gene variants on chromosome 4q25 and atrial fibrillation.

Phenotype and genotype GWAS data, including those from the Framingham Heart Study, made available in the NIH Database of Genotype and Phenotype (dbGaP)

Phenotype and genotype GWAS data made available in the NIH dbGaP for the MultiEthnic Study of Atherosclerosis and the Women's Health Initiative

2008 Reports of multiple gene loci associated with lipid traits and with body-mass index using GWAS

Reports of loci associated with intracranial aneurysm

Reports of cardiovascular pharmacogenetic effects (prolonged QT interval, increased warfarin maintenance dose, statin-induced myopathy) using GWAS

NHLBI begins whole-exome and whole-genome sequencing of the Framingham Heart Study and other cohorts for studying cardiovascular disease. **2009**

GWAS implicate genetic variants in early-onset myocardial infarction, QT-interval prolongation, blood pressure and hypertension, echocardiographic measures of cardiac structure, and blood-cell traits.

Exome sequencing identifies mutations in *SLC26A3* for renal salt wasting in Bartter's syndrome.

2010 Large GWAS for CVD risk factors report total of 95 gene loci associated with lipid traits, >30 gene loci associated with type 2 diabetes mellitus, >30 gene loci associated with body-mass index and obesity, and hundreds of variants associated with height.

GWAS uncover gene loci associated with cigarette-smoking behavior, sudden cardiac death and ventricular fibrillation, thrombosis factors and platelet function, and ECG interval (RR, PR, QRS).

Variants in 9p21 are found to be associated with abdominal aortic aneurysm.

Exome sequencing identifies *ANGPTL3* mutations in familial combined hypolipidemia.

Whole-genome sequencing identifies *ABCG5* mutations in severe hypercholesterolemia.

Initial report of genomewide sequencing of multiple persons in the 1000 Genomes project

Large GWAS yields >30 gene loci associated with clinical coronary artery disease. **2011**

Large GWAS reports >30 gene loci associated with blood pressure and hypertension.

GWAS also yield variants in *MYH6* associated with sick sinus syndrome and variants associated with dilated cardiomyopath and carotid intima–media thickness.

Figure 1. Timeline of Genetic and Genomic Research in Cardiovascular Medicine.

The New England Journal of Medicine
Downloaded from nejm.org at VIRGINIA COMMONWEALTH UNIV on December 1, 2011. For personal use only. No other uses without permission.
Copyright © 2011 Massachusetts Medical Society. All rights reserved.

ponents of cardiovascular risk factors and subclinical disease, including coronary artery disease.[10] Approaches to identifying the genetic causes of polygenic cardiovascular diseases (and other polygenic diseases) before completion of the draft sequence of the human genome were largely unsuccessful. A decade later, hundreds of loci associated with many cardiovascular diseases and traits have been identified. This genetic bounty is the yield of genomewide association studies, which involve testing of a large set of genetic variants in case and control subjects from a population to determine which variants are associated with the disease in question (see the Glossary).[11]

The common disease–common variant hypothesis proposes that common variants, which are defined as variants with a prevalence of at least 5% in the population, have a role in the cause and

Glossary

Allele: One of two or more copies of a genetic sequence at a chromosomal location. Alleles can be considered according to their frequencies in the human population, ranging from common variants (minor allele frequency, >5%) to low-frequency variants (minor allele frequency, 0.5 to 5%) to rare or private alleles in one or a limited number of families (minor allele frequency, <0.5%). A null allele is not functional.

Candidate-gene study: An approach used in genetics research that focuses on suspected, or candidate, genes that have been selected because of a perceived match between their known or presumed function and biologic function in the disease under investigation.

Exome: All known protein-coding sequences, or exons, in the human genome, constituting approximately 1 to 2% of the 3.2 billion nucleotide base pairs in the human genome.

Gene enhancer: A short region of DNA that can be bound with proteins, such as trans-acting factors (i.e., factors acting from a different molecule), to enhance transcription levels of genes in a gene cluster. Although enhancers are usually cis-acting (i.e., acting from the same molecule), an enhancer does not need to be close to the genes it acts on and may not be located on the same chromosome.

Genomewide association study: An approach used in genetics research to look for associations between many (typically hundreds of thousands) specific genetic variations (most commonly single-nucleotide polymorphisms) and particular diseases.

Genotyping array: A technique used to study many genes at once. Thousands of gene sequences are placed in known locations on a glass slide. A sample containing DNA or RNA is deposited on the slide, or gene chip. The binding of complementary base pairs from the sample and the gene sequences on the chip can be measured with the use of fluorescence to detect the presence and determine the amount of specific sequences in the sample.

Induced pluripotent stem cell: A type of pluripotent stem cell derived from a nonpluripotent cell, typically an adult somatic cell such as a fibroblast, by transfection of a stem-cell–associated gene.

Haplotype maps: Maps of a set of DNA variations, or polymorphisms, that tend to be inherited together. Haplotype maps are used to aid gene-discovery efforts.

Haplotype structure: Patterns of haplotypes that define the structure of regions of the genome. Haplotype structure is used to compare genome regions across populations or species.

Locus: The specific chromosomal location of a gene or other DNA sequence of interest.

Mendelian inheritance: Patterns of inheritance characteristic of organisms that reproduce sexually, as described by Austrian monk Gregor Mendel in the mid-19th century.

Mendelian randomization: A method that uses measured variation in genes of known function to examine the causal effect of a modifiable exposure on a disease. It provides unbiased estimates of the effects of a putative casual variable without the need for a traditional randomized trial.

Next-generation sequencing: DNA sequencing that harnesses advances in miniaturization technology to simultaneously sequence multiple areas of the genome rapidly and at low cost.

Noncoding DNA sequence: A DNA sequence that does not encode proteins. Noncoding DNA sequences, once referred to as "junk DNA," account for the majority of genome sequences and are now known to harbor regions that regulate gene expression.

Polygenic: Produced by two or more genes.

Single-nucleotide polymorphism: A single-nucleotide variation in a genetic sequence; a common form of variation in the human genome.

Splice variants: Abnormal variations in RNA splicing that are implicated in a disease. Many genetic disorders result from splicing variants.

Whole-exome sequencing: Sequencing of the coding regions, or exons, of an entire genome.

The New England Journal of Medicine
Downloaded from nejm.org at VIRGINIA COMMONWEALTH UNIV on December 1, 2011. For personal use only. No other uses without permission.
Copyright © 2011 Massachusetts Medical Society. All rights reserved.

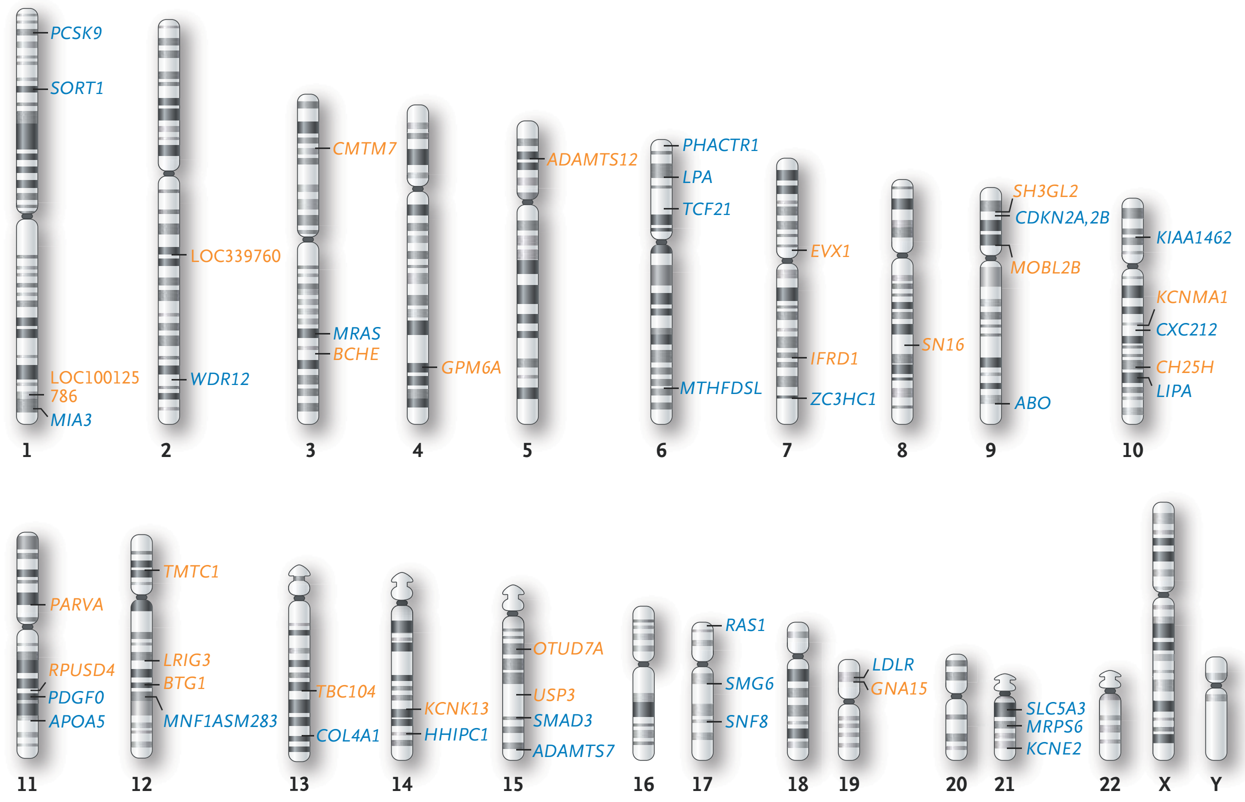

PCSK9
SORT1
LOC100125
786
MIA3
1
LOC339760
WDR12
2
CMTM7
MRAS
BCHE
3
GPM6A
4
ADAMTS12
5
PHACTR1
LPA
TCF21
MTHFDSL
6
EVX1
IFRD1
ZC3HC1
7
SN16
8
SH3GL2
CDKN2A,2B
MOBL2B
ABO
9
KIAA1462
KCNMA1
CXC212
CH25H
LIPA
10
PARVA
RPUSD4
PDGF0
APOA5
11
TMTC1
LRIG3
BTG1
MNF1ASM283
12
TBC104
COL4A1
13
KCNK13
HHIPC1
14
OTUD7A
USP3
SMAD3
ADAMTS7
15
16
RAS1
SMG6
SNF8
17
18
LDLR
GNA15
19
20
SLC5A3
MRPS6
KCNE2
21
22
X
Y

Figure 2. Genomic Locations of Genetic Variants Associated with the Risk of Myocardial Infarction and Heart Failure.

Genetic variants that are associated with myocardial infarction (blue) and heart failure (orange) are shown according to their chromosomal location. LOC denotes locus.

pathophysiology of common diseases. It is on this premise that the genomewide association study is based, since it comprises tests of association between disease and common variants spread throughout the genome. On the basis of this approach, researchers have assembled catalogues of cardiovascular variants, using genotyping arrays, haplotype maps, and statistical methods. Changes in data-sharing policies have led to the creation of publicly available genome databases. International collaborations combining study cohorts, often including tens of thousands of research participants, have been formed.[12]

The astonishingly large number of new loci associated with cardiovascular risk factors, subclinical indexes, and disease end points have provided insights into the biologic pathways that underlie disease (Fig. 2). The application of such findings to the prediction of risk and to the prevention and treatment of disease is premature and awaits considerable research.

CORONARY ARTERY DISEASE AND MYOCARDIAL INFARCTION

Genomewide association studies have identified about 30 loci associated with myocardial infarction and coronary artery disease (Table 1).[13-35] A meta-analysis of 14 such studies that involved 22,233 case subjects with coronary artery disease and 64,762 control subjects of European descent and that were followed by replication studies involving 56,682 case and control subjects identified 13 new loci associated with coronary artery disease, in addition to confirming 10 of 12 previously reported loci.[14] *ABO* and *ADAMTS7* were found to be associated with angiographically confirmed coronary atherosclerosis, *CNNM2* with high blood pressure, and the *APOA5* gene cluster with elevated levels of triglycerides and cholesterol subfractions. The majority of loci associated with myocardial infarction reside in genomic regions that have not previously been implicated in coronary artery disease; only a minority of loci mediate an

The New England Journal of Medicine
Downloaded from nejm.org at VIRGINIA COMMONWEALTH UNIV on December 1, 2011. For personal use only. No other uses without permission.
Copyright © 2011 Massachusetts Medical Society. All rights reserved.

Table 1. Representative Genomewide Association Studies of Common Cardiovascular Diseases.

Study	Clinical Outcome	Sample Size		Major Ethnic Group	Selected Major Findings
		Genomewide Association	Replication		
		number of subjects			
Kathiresan et al., 2009[13]	Myocardial infarction	6,042	19,492	European	9 Loci, including 9p21 and *SORT1*
Schunkert et al., 2011[14]	Coronary artery disease	86,995	56,682	European	>20 Loci, including *ABO* and *ADAMTS7*
C4D Consortium, 2011[15]	Coronary artery disease	30,482	40,593	European, East Asian	Novel loci, including *LIPA*
Smith et al., 2010[16]	Heart failure incidence	23,821	NA	European, African	Suggestive loci, including *LRIG3*
Morrison et al., 2010[17]	Heart failure mortality	2,992	NA	European, African	Suggestive loci, including *ADAMTS12*
Villard et al., 2011[18]	Dilated cardiomyopathy	2,287	2,467	European	Top loci: *HSPB7* and *BAG3*
Ikram et al., 2009[19]	Ischemic stroke	19,602	7,269	European, African	Top locus: *NINJ2*
Yamada et al., 2009[20]	Ischemic stroke	267	5,981	Japanese	Top locus: *CELSR1*
Akiyama et al., 2010[21]	Intracranial aneurysm	482	1,398	Japanese	Suggestive loci, including *ARHGEF11* and *TMEM195*
Bilguvar et al., 2008[22]	Intracranial aneurysm	7,856	1,171	European, Japanese	Top loci: 9p21 and *SOX17*
Yasuno et al., 2010[23]	Intracranial aneurysm	15,295	4,777	European, Japanese	Associations with 9p21 and *CNNM2*
Koriyama et al., 2010[24]	Peripheral arterial disease	1,553	2,239	Japanese	Top loci: *OSBPL10* and 10p12.31
Thorgeirsson et al., 2008[25]	Peripheral arterial disease	10,995	4,848	European	Top locus: *CHRNA3*
Elmore et al., 2009[26]	Abdominal aortic aneurysm	235	1,421		Suggestive locus: 3p12.3
Gretarsdottir et al., 2010[27]	Abdominal aortic aneurysm	31,795	10,718	European	Top loci: 9p21 and *DAB2IP*
Benjamin et al., 2009[28]	Lone atrial fibrillation	40,518	6,218	European	Top loci: *PITX2* and *ZFHX3*
Gudbjartsson et al., 2007[29]	Lone atrial fibrillation	36,137	5,806	European	Top loci: 4q25 and *ZFHX3*
Ellinor et al., 2010[30]	Lone atrial fibrillation	14,179	4,771	European	Top loci: *KCNN3* and 20q13.13
Holm et al., 2011[31]	Sick sinus syndrome	38,384	1,654	European	Top locus: *MYH6*
Bezzina et al., 2010[32]	Ventricular fibrillation	972	537	European	Top locus: *CXADR*
Arking et al., 2010[33]	Sudden cardiac arrest	650	19,611	European	Top locus: *GPC5*
Arking et al., 2011[34]	Sudden cardiac death	22,055	14,265	European	Top locus: *BAZ2B*
Trégouët et al., 2009[35]	Venous thromboembolism	1,647	3,237	European	Top locus: *AB0*

effect through known risk factors. A second meta-analysis of genomewide association studies involving more than 30,000 case and control subjects showed an additional four new loci associated with coronary artery disease in multiple ethnic groups.[15]

Studies of early-onset myocardial infarction have identified more than 10 risk loci.[13] The most strongly associated locus is 9p21.[36,37] This region harbors genes (*CDKN2A* and *CDKN2B*) that are implicated in cell cycling and cancer, although the single-nucleotide polymorphisms (SNPs) associated with myocardial infarction are located not near these or other protein-coding genes but rather within a noncoding RNA molecule called ANRIL (for antisense noncoding RNA in the *INK4* locus). The expression of ANRIL splice variants, but not *CDKN2B* or other nearby genes, is associated with atherosclerosis.[38] SNPs at the 9p21 locus are also associated with other cardiovascular diseases, including stroke[39] and aortic aneurysm.[40]

Loci associated with coronary artery disease harbor genes known to be important in lipid variation, including *SORT1*, *PCSK9*, *HNF1A*, *MRAS*, and *LPA*. The position of other SNPs implicates inflam-

The New England Journal of Medicine
Downloaded from nejm.org at VIRGINIA COMMONWEALTH UNIV on December 1, 2011. For personal use only. No other uses without permission.
Copyright © 2011 Massachusetts Medical Society. All rights reserved.

matory processes in conferring a risk of coronary atherosclerosis.

HEART FAILURE

Genomewide association studies have identified many possible loci associated with heart failure and death from heart failure, although few of such studies have been replicated.[16,17] Genomewide association studies for heart failure have been limited by modest numbers of cases of heart failure (relative to the number of such studies for coronary artery disease) and the heterogeneous nature (and thus heterogeneic sets of cases) of heart failure. A recent genomewide association study for idiopathic dilated cardiomyopathy identified variants in *HSPB7*, encoding a heat-shock protein previously implicated in heart failure, and *BAG3*; a marked myopathy develops in mice deficient in *Bag3*.[18]

ARRHYTHMIAS

Genomewide association studies have uncovered genetic variants for arrhythmias, including atrial fibrillation,[28-30] ventricular fibrillation,[32] sudden cardiac death,[33,34] and the sick sinus syndrome[31] (Table 1). *MYH6*, a previously unidentified gene associated with susceptibility to the sick sinus syndrome, encodes the alpha heavy-chain subunit of cardiac myosin,[31] suggesting that myosin proteins may regulate cardiac conduction in addition to myocyte function.

PERIPHERAL AND CEREBRAL VASCULAR DISEASES

Genomewide association studies have yielded evidence of varying strength for genetic loci associated with ischemic stroke,[19,20] intracranial aneurysm,[21-23] peripheral arterial disease,[24,25] aortic aneurysm,[25-27] venous thromboembolism,[35] and erythrocyte phenotypes[41] (Table 1). In some cases, these loci are common to coronary artery disease and myocardial infarction, suggesting a common genetic contribution to multiple vascular beds.

MODIFIABLE RISK FACTORS AND SUBCLINICAL DISEASE

Genome consortia with sample sizes exceeding 10,000 cases have investigated major modifiable risk factors (e.g., hypertension, dyslipidemia, type 2 diabetes, and cigarette smoking) and obesity, using quantitative measures and clinically relevant extreme end points (Table 2). These studies point to loci associated with hypertension (as defined by the use of antihypertensive therapy or elevations in blood pressure),[42] LDL and high-density lipoprotein (HDL) cholesterol and triglycerides,[43] the number of cigarettes smoked,[44] type 2 diabetes (with the use of quantitative traits of fasting glucose and glycated hemoglobin levels),[45,46] and obesity (with the use of body-mass index and adiposity-related traits).[47,48] Studies have also identified multiple loci associated with other presumed quantitative biomarkers of cardiovascular risk and presymptomatic coronary artery disease (e.g., fibrinogen,[49] C-reactive protein,[50] intercellular adhesion molecule 1,[51] plasma homocysteine,[52] and carotid-artery intima–media thickness and plaque[53]).

Newer approaches for establishing evidence of causality use human populations as a model system. Mendelian randomization analysis takes advantage of the lifelong association between a risk allele and a quantitative measure of a biomarker for clinical coronary artery disease to estimate whether there is evidence of a causal association between the biomarker and the disease. Mendelian randomization studies have used genetic variants that have previously been implicated by genomewide association studies to provide evidence for a causal association between LDL cholesterol (variants in *LDLR*)[54] or lipoprotein(a) (variants in *LPA*)[55] and coronary artery disease and evidence against a causal association between C-reactive protein (variants in *CRP*) and coronary artery disease.[56]

Current genomewide association studies focus on large populations in order to strengthen the evidence for association and replication. For example, a genomewide association study of more than 100,000 research participants identified 95 loci associated with at least one of three lipids: LDL and HDL cholesterol and triglycerides.[43] Although each individual variant had only a modest effect, the combined effect of the 95 loci explained approximately 25% of the genetic variance in LDL and HDL cholesterol levels. Similarly, large meta-analyses of variants that are associated with type 2 diabetes[45] and myocardial infarction[14] suggest that these variants, in aggregate, account for 25% and 10%, respectively, of the inherited variation in disease outcomes.

We have learned several lessons from genomewide association studies. Although some risk loci that have been discovered through this approach encompass genes that had been previously impli-

The New England Journal of Medicine
Downloaded from nejm.org at VIRGINIA COMMONWEALTH UNIV on December 1, 2011. For personal use only. No other uses without permission.
Copyright © 2011 Massachusetts Medical Society. All rights reserved.

Table 2. Representative Large-Scale Genomewide Association Studies of Risk Factors for Cardiovascular Disease.*

Study	Risk Factor or Clinical Trait	Sample Size		Major Ethnic Group	Selected Major Findings
		Genomewide Association	Replication		
		number of subjects			
Ehret et al., 2011[42]	Systolic and diastolic blood pressure, hypertension	69,395	133,661	European	>25 Loci ,including *CACNB2* and *SH2B3*
Teslovich et al., 2010[43]	Total and LDL cholesterol	100,184	39,875	European, South Asian, East Asian, African	>35 Loci, including *SORT1* and *HMGCR*
Teslovich et al., 2010[43]	HDL cholesterol	100,184	39,875	European, South Asian, East Asian, African	>35 Loci, including *SCARB1* and *CETP*
Teslovich et al., 2010[43]	Triglycerides	100,184	39,875	European, South Asian, East Asian, African	>20 Loci, including *ANGPTL3* and *JMJD1C*
Thorgeirsson et al., 2008[25]	Quantity of cigarettes smoked	31,266	54,731	European	Top loci: *CHRNB3* and 15q25
TAGC, 2010[44]	Quantity of cigarettes smoked	74,053	68,988	European	Top loci: *DBH* and *CYP2A6*
Voight et al., 2010[45]	Type 2 diabetes mellitus	47,117	94,337	European	>25 Loci, including *TCF7L* and *IRS1*
Dupuis et al., 2010[46]	Fasting glucose level	46,186	76,558	European	>20 Loci, including *GCKR* and *ADRA2A*
Speliotes et al., 2010[47]	Body-mass index	123,865	125,931	European	>30 Loci, including *FTO* and *TMEM18*
Heard-Costa et al., 2009[48]	Waist circumference	31,373	38,641	European	Top loci: *NRXN2* and *MC4R*

* Genomewide association studies were considered to be large in scale if they include a sample of more than 10,000 subjects. HDL denotes high-density lipoprotein, LDL low-density lipoprotein, and TAGC Tobacco and Genetics Consortium.

cated, most of such loci implicate genes that had not been thought to have a role in conferring disease risk. Many loci show association with cardiovascular disease across groups of differing ancestry, and most cardiovascular traits are influenced by a large number of loci. However, limitations of genomewide association studies have prevented immediate translation of these findings into clinical practice, since each variant has a very small effect and is therefore not useful for prediction. Moreover, the implicated variants are rarely themselves the causal variants; rather, they are linked to the true causal variants, and identification of the latter usually warrants a great deal of additional work.

TARGETED AND GENOMEWIDE DNA SEQUENCING

The next generation of genomic approaches is upon us, thanks to the increasing efficiency and decreasing cost of sequencing technology. Deep DNA sequencing is performed with the use of miniaturized technology that simultaneously sequences multiple areas of the genome. This approach is used to sequence candidate regions and specific components of the human genome, such as exons or noncoding DNA. Sequencing studies in community-based cohorts have shown that at least 1 of every 64 persons carries a functional mutation in one of three genes (*NCCT*, *NKCC2*, or *ROMK*) that is associated with clinically significant alterations in blood pressure[57] and that there is a striking excess of nonsynonymous variants in the gene encoding adipokine (*ANGPTL4*) in persons with low triglyceride levels.[58] Targeted sequencing in a population with extreme values of LDL cholesterol identified variants in *PCSK9* that occur in up to 3% of the population. Such variants are associated with a low LDL cholesterol level and a decreased risk of incident coronary artery disease.[59] A sequencing study of *FAAH* and *MGLL* at extremes of the body-mass index uncovered several rare variants in promoter and enhancer regions of these genes, suggesting that sequence variation may regulate extreme obesity.[60]

Target genes for sequencing also include those identified by genomewide association studies that

The New England Journal of Medicine
Downloaded from nejm.org at VIRGINIA COMMONWEALTH UNIV on December 1, 2011. For personal use only. No other uses without permission.
Copyright © 2011 Massachusetts Medical Society. All rights reserved.

are associated with cardiovascular disease and its risk factors. Resequencing of exons in *APOA5*, *GCKR*, *LPL*, and *APOB* — genes that are implicated in hypertriglyceridemia — revealed twice as many rare missense or nonsense variants in persons with high triglyceride levels than in control subjects, corresponding to a carrier frequency of 28.1% in affected persons and 15.3% in control subjects.[61]

Whole-exome sequencing has uncovered genetic variants in persons with rare forms of diseases that affect blood pressure and circulating blood lipids, including a missense mutation in *SLC26A3* (a congenital chloride diarrhea locus) in a patient with a suspected diagnosis of the renal salt-wasting disease Bartter's syndrome,[62] a variant in *ANGPTL3* in a family with familial combined hypolipidemia,[63] and variants in *BAG3* in families with familial dilated cardiomyopathy.[64] Whole-exome sequencing in large populations may help to identify rare genetic variants in relatively common cardiovascular diseases, such as early myocardial infarction, that confer a major risk of complications and death and for which there is good evidence of heritability; indeed, this is a major focus of whole-exome sequencing at present. Whole-genome sequencing has also confirmed single-gene disease variants in families with rare diseases,[65] including mutations in *ABCG* in an infant with severe hypercholesterolemia.[66]

CARDIOVASCULAR DISEASE PATHWAYS

The best yield of genomewide association studies is the provision of insights into the biologic pathways — often previously unsuspected — that underlie causes of disease. Such insights have led to hypothesis-driven investigations of implicated pathways with the use of molecular, genetic, biochemical, and cellular approaches. The genetic control of myocardial infarction and lipids provides a case in point.

The initial findings that connected genetic variation in the 9p21 region with atherosclerotic progression and myocardial infarction caught cardiovascular scientists by surprise, because this region is devoid of genes that had previously been associated with coronary artery disease. Subsequent research has revealed possible mechanisms by which genes in this region may contribute to atherosclerosis. A recent report implicated the 9p21 risk interval in regulating the expression of cardiac *CDKN2A/B* expression genes.[38] Another report implicated inflammatory pathways through 33 gene enhancers located in the 9p21 region.[67] Two SNPs located in one of these enhancers disrupt a binding site for STAT1, a signal transduction protein that regulates inflammation. This enhancer locus physically interacts with the *CDKN2A/B* locus and an interval downstream of *IFNA21*, the gene that encodes interferon-γ in human vascular endothelial cells. The activation of interferon-γ affects transcriptional regulation of the 9p21 locus, including STAT1 binding, suggesting a link between genetic susceptibility to coronary artery disease and response to inflammatory signaling in vascular cells. Yet another driver of disease may be the aforementioned noncoding RNA element ANRIL.

Loci that are discovered by genomic approaches are also relevant to lipid biology and potential therapeutic targets. The gene encoding sortilin (*SORT1*) contains a common variant that creates a binding site for a transcription factor that when bound to this site alters liver expression of LDL cholesterol in humans. Transgenic studies have shown that mice with *Sort1* that contains the transcription-factor binding site have altered plasma levels of LDL cholesterol, suggesting a previously unknown regulatory pathway for LDL cholesterol.[68] A potentially therapeutic target was implicated by the discovery of a pair of common nonsense variants in *PCSK9* in patients of African descent. An estimated 2.6% of persons of African descent carry one of these variants, and each variant results in comparatively low lipid levels and a reduced susceptibility to myocardial infarction.[69] Humans who are homozygous for this null allele are healthy and have a reduced risk of myocardial infarction, suggesting that the absence of PCSK9 may be well tolerated and thus rendering PCSK9 an attractive drug target.

SNPs that more recently have been found to be associated with cardiovascular disease and its risk factors implicate components of pathways previously identified with risk factors for or mechanisms of the disease. For example, the 95 loci associated with levels of LDL and HDL cholesterol or triglycerides in more than 100,000 persons of European ancestry implicate nearly all the 18 genes that have previously been shown to be mutated

The New England Journal of Medicine
Downloaded from nejm.org at VIRGINIA COMMONWEALTH UNIV on December 1, 2011. For personal use only. No other uses without permission.
Copyright © 2011 Massachusetts Medical Society. All rights reserved.

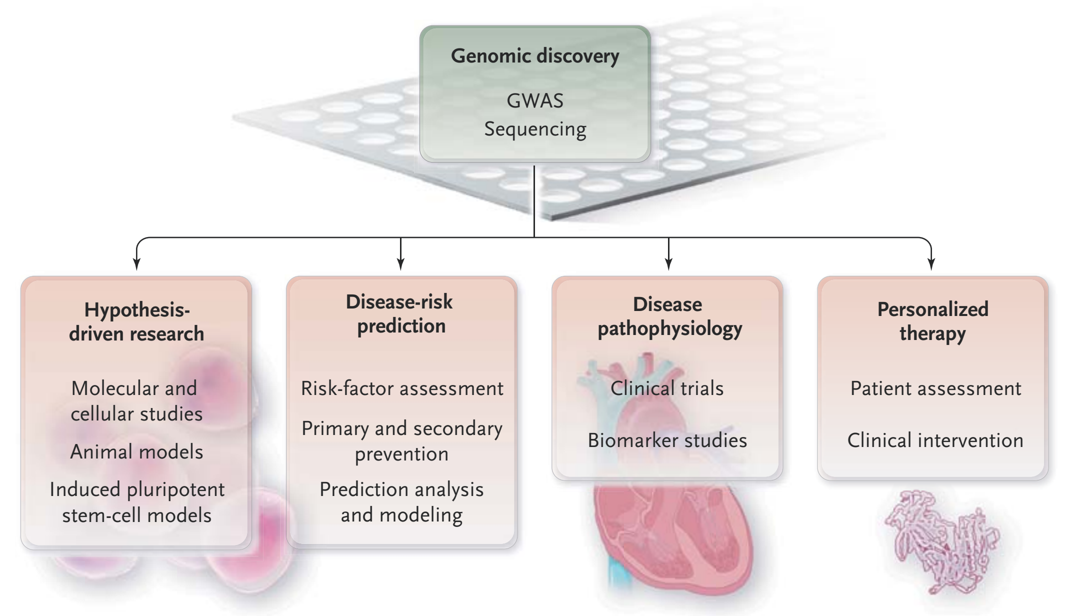

Genomic discovery
GWAS
Sequencing
Hypothesis-driven research
Molecular and cellular studies
Animal models
Induced pluripotent stem-cell models
Disease-risk prediction
Risk-factor assessment
Primary and secondary prevention
Prediction analysis and modeling
Disease pathophysiology
Clinical trials
Biomarker studies
Personalized therapy
Patient assessment
Clinical intervention

Figure 3. Investigative Pathways Leading from Gene Discovery to Clinical Application.

Genomewide association studies (GWAS) and gene sequencing represent discovery research. Findings from these studies stimulate further investigations in hypothesis-driven research and disease-risk prediction. Applications to clinical cardiovascular care will follow.

in rare mendelian lipid disorders.[43] Most of these loci were also associated with cardiovascular disease in persons of African descent and in persons of Asian descent. Some loci housed both common and rare variants; in such cases, the common variants explained more of the heritability.

An exciting approach is the use of induced pluripotent stem cells to model cardiovascular disease and test hypotheses generated from genomic studies. Two lines of induced pluripotent stem cells, which were generated from dermal fibroblasts in patients who had the long-QT syndrome with specific genetic mutations, have been differentiated into cardiomyocytes and are being used to investigate electrophysiological properties, such as action potentials and ion fluxes.[70,71]

PREDICTION, PREVENTION, AND TREATMENT

The improved understanding of cardiovascular pathophysiology that has been achieved through genetic discovery provides new opportunities for prediction, prevention, and treatment (Fig. 3). Genetic risk prediction is at an early stage, and insufficient evidence exists at present to warrant the use of a genetic risk score on the basis of SNPs identified through genomic approaches.[72,73] Additional research is needed to prospectively assess the utility of genetic risk scores in the prediction of cardiovascular disease, such as myocardial infarction and coronary artery disease, before clinical use. Some observers have suggested that 150 genes with odds ratios of 1.5 or 250 genes with odds ratios of 1.25 will be needed.[74] Future studies of genetic risk scores will probably require hundreds of associated SNPs (identified through genomewide association studies), combined with low-frequency risk alleles discovered through whole-genome sequencing, to provide evidence-based prediction.

Studies are under way to investigate the use of next-generation sequencing for the screening of rare forms of cardiovascular disease[75] and to annotate all mutations within an individual's genome.[76] Although the costs of accurate whole-genome sequencing are dropping dramatically, the costs of data analysis and storage remain high and are barriers to clinical implementation.

Loci that are uncovered by genomewide association studies, despite their modest effects, may

The New England Journal of Medicine
Downloaded from nejm.org at VIRGINIA COMMONWEALTH UNIV on December 1, 2011. For personal use only. No other uses without permission.
Copyright © 2011 Massachusetts Medical Society. All rights reserved.

have therapeutic implications. For example, a common variant at the *HMGCR* locus (with a prevalence of 40% among persons of European ancestry) is associated with a very modest elevation in the LDL cholesterol level (2.8 mg per deciliter [0.07 mmol per liter]).[43] There are no known rare mutations at this locus, presumably because they would be lethal. Yet the encoded protein, 3-hydroxy-3-methylglutaryl coenzyme A (HMG-CoA) reductase, is the target of statin drugs, which reduce LDL cholesterol levels and the risk of myocardial infarction.

Pharmacogenomics has potential near-term application. Discoveries that are provided by genomewide association studies have strengthened the evidence for pharmacogenetic interactions in a number of commonly used cardiovascular drugs. Variants in *CYP2C9* and *VKORC1* explain up to 40% of the variation in the adjusted dose of warfarin, and the Food and Drug Administration (FDA) recently revised the label on warfarin to allow for genotype-specific dose ranges.[77] Variations in a cytochrome P-450 enzyme, *CYP2C19*, are associated with decreased antiplatelet efficacy and an increased risk of cardiovascular disease among patients taking the antiplatelet drug clopidogrel, which prompted an FDA warning and recommendations for a dose adjustment or use of alternative drugs in patients with this variant.[77]

Evidence of other pharmacogenetic interactions that may be clinically important is accumulating. Variation in the β_1-adrenergic–receptor gene, *ADRB1*, is associated with altered responsiveness to beta-blockade in heart failure.[78] A variant in *SLC01B1* has been implicated in statin-related myopathy.[79] Deep sequencing of variants and genes related to drug absorption, distribution, metabolism, and excretion may identify specific variants that contribute to the heterogeneity of responsiveness to cardiovascular drugs.

FUTURE DIRECTIONS

The field of cardiovascular genomics has two distinct goals: understanding biologic mechanisms and applying that knowledge to personalized medicine. Knowledge of molecular pathways can lead to improved therapeutics on a broad basis (regardless of the individual genotype) or at an individualized level (targeted specifically to the genotype). During the past 5 years, the discovery of hundreds of cardiovascular loci is a start. In the years to come, we will require studies of tens of thousands of patients with cardiovascular disease that combine tests of genomewide association (involving missense, rare, and common variants) with sequencing.[80] Functional studies are essential to characterize molecular and cellular pathways and to develop appropriately targeted therapies. Genomics is permeating biomedical research, and these genomic advances must proceed through basic discoveries, functional characterization, preclinical proof-of-principle studies, first-in-human studies, clinical trials, and regulatory approval. Cardiovascular science and medicine have made enormous strides over the past century, beginning with the brilliant elucidation of cardiovascular physiology and leading to molecular and cellular studies, with concurrent epidemiologic determinations of risk factors. The cardiovascular field is now primed for genomic medicine to make equal, if not greater, contributions.

Disclosure forms provided by the authors are available with the full text of this article at NEJM.org.

REFERENCES

1. NHLBI fact book, fiscal year 2009. Bethesda, MD: National Heart, Lung, and Blood Institute, February 2010 (http://www.nih.gov/about/FactBook2009_final.pdf).

2. Kucharska-Newton AM, Couper DJ, Pankow JS, et al. Hemostasis, inflammation, and fatal and nonfatal coronary heart disease: long-term follow-up of the Atherosclerosis Risk in Communities (ARIC) cohort. Arterioscler Thromb Vasc Biol 2009; 29:2182-90.

3. Kathiresan S, Larson MG, Keyes MJ, et al. Assessment by cardiovascular magnetic resonance, electron beam computed tomography, and carotid ultrasonography of the distribution of subclinical atherosclerosis across Framingham risk strata. Am J Cardiol 2007;99:310-4.

4. Greenland P, Alpert JS, Beller GA, et al. 2010 ACCF/AHA guideline for assessment of cardiovascular risk in asymptomatic adults: a report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines. Circulation 2010;122(25):e584-e636.

5. Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. Nature 2001;409:860-921. [Errata, Nature 2001;411:720, 412:565.]

6. Venter JC, Adams MD, Myers EW, et al. The sequence of the human genome. Science 2001;291:1304-51.

7. Nabel EG. Cardiovascular disease. N Engl J Med 2003;349:60-72. [Erratum, N Engl J Med 2003;349:620.]

8. Goldstein JL, Brown MS. The LDL receptor. Arterioscler Thromb Vasc Biol 2009; 29:431-8.

9. Lee DS, Pencina MJ, Benjamin EJ, et al. Association of parental heart failure with risk of heart failure in offspring. N Engl J Med 2006;355:138-47.

10. Parikh NI, Hwang SJ, Larson MG, et al. Parental occurrence of premature cardio-

The New England Journal of Medicine
Downloaded from nejm.org at VIRGINIA COMMONWEALTH UNIV on December 1, 2011. For personal use only. No other uses without permission.
Copyright © 2011 Massachusetts Medical Society. All rights reserved.

vascular disease predicts increased coronary artery and abdominal aortic calcification in the Framingham Offspring and Third Generation cohorts. Circulation 2007;116:1473-81.

11. Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. Nat Genet 2003;33:177-82.

12. O'Donnell CJ, Nabel EG. Cardiovascular genomics, personalized medicine, and the National Heart, Lung, and Blood Institute: part I: the beginning of an era. Circ Cardiovasc Genet 2008;1:51-7.

13. Kathiresan S, Voight BF, Purcell S, et al. Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. Nat Genet 2009;41:334-41. [Erratum, Nat Genet 2009;41:762.]

14. Schunkert H, Konig IR, Kathiresan S, et al. Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. Nat Genet 2011;43:333-8.

15. Coronary Artery Disease (C4D) Genetics Consortium. A genome-wide association study in Europeans and South Asians identifies five new loci for coronary artery disease. Nat Genet 2011;43:339-44.

16. Smith NL, Felix JF, Morrison AC, et al. Association of genome-wide variation with the risk of incident heart failure in adults of European and African ancestry: a prospective meta-analysis from the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) consortium. Circ Cardiovasc Genet 2010;3:256-66.

17. Morrison AC, Felix JF, Cupples LA, et al. Genomic variation associated with mortality among adults of European and African ancestry with heart failure: the cohorts for heart and aging research in genomic epidemiology consortium. Circ Cardiovasc Genet 2010;3:248-55.

18. Villard E, Perret C, Gary F, et al. A genome-wide association study identifies two loci associated with heart failure due to dilated cardiomyopathy. Eur Heart J 2011;32:1065-76.

19. Ikram MA, Seshadri S, Bis JC, et al. Genomewide association studies of stroke. N Engl J Med 2009;360:1718-28.

20. Yamada Y, Fuku N, Tanaka M, et al. Identification of CELSR1 as a susceptibility gene for ischemic stroke in Japanese individuals by a genome-wide association study. Atherosclerosis 2009;207:144-9.

21. Akiyama K, Narita A, Nakaoka H, et al. Genome-wide association study to identify genetic variants present in Japanese patients harboring intracranial aneurysms. J Hum Genet 2010;55:656-61.

22. Bilguvar K, Yasuno K, Niemelä M, et al. Susceptibility loci for intracranial aneurysm in European and Japanese populations. Nat Genet 2008;40:1472-7.

23. Yasuno K, Bilguvar K, Bijlenga P, et al. Genome-wide association study of intracranial aneurysm identifies three new risk loci. Nat Genet 2010;42:420-5.

24. Koriyama H, Nakagami H, Katsuya T, et al. Identification of evidence suggestive of an association with peripheral arterial disease at the OSBPL10 locus by genome-wide investigation in the Japanese population. J Atheroscler Thromb 2010;17:1054-62.

25. Thorgeirsson TE, Geller F, Sulem P, et al. A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. Nature 2008;452:638-42.

26. Elmore JR, Obmann MA, Kuivaniemi H, et al. Identification of a genetic variant associated with abdominal aortic aneurysms on chromosome 3p12.3 by genome wide association. J Vasc Surg 2009;49:1525-31.

27. Gretarsdottir S, Baas AF, Thorleifsson G, et al. Genome-wide association study identifies a sequence variant within the DAB2IP gene conferring susceptibility to abdominal aortic aneurysm. Nat Genet 2010;42:692-7.

28. Benjamin EJ, Rice KM, Arking DE, et al. Variants in ZFHX3 are associated with atrial fibrillation in individuals of European ancestry. Nat Genet 2009;41:879-81.

29. Gudbjartsson DF, Arnar DO, Helgadottir A, et al. Variants conferring risk of atrial fibrillation on chromosome 4q25. Nature 2007;448:353-7.

30. Ellinor PT, Lunetta KL, Glazer NL, et al. Common variants in KCNN3 are associated with lone atrial fibrillation. Nat Genet 2010;42:240-4.

31. Holm H, Gudbjartsson DF, Sulem P, et al. A rare variant in MYH6 is associated with high risk of sick sinus syndrome. Nat Genet 2011;43:316-20.

32. Bezzina CR, Pazoki R, Bardai A, et al. Genome-wide association study identifies a susceptibility locus at 21q21 for ventricular fibrillation in acute myocardial infarction. Nat Genet 2010;42:688-91.

33. Arking DE, Reinier K, Post W, et al. Genome-wide association study identifies GPC5 as a novel genetic locus protective against sudden cardiac arrest. PLoS ONE 2010;5(3):e9879.

34. Arking DE, Junttila MJ, Goyette P, et al. Identification of a sudden cardiac death susceptibility locus at 2q24.2 through genome-wide association in European ancestry individuals. PLoS Genet 2011;7(6):e1002158.

35. Trégouët DA, Heath S, Saut N, et al. Common susceptibility alleles are unlikely to contribute as strongly as the FV and ABO loci to VTE risk: results from a GWAS approach. Blood 2009;113:5298-303.

36. McPherson R, Pertsemlidis A, Kavaslar N, et al. A common allele on chromosome 9 associated with coronary heart disease. Science 2007;316:1488-91.

37. Helgadottir A, Thorleifsson G, Manolescu A, et al. A common variant on chromosome 9p21 affects the risk of myocardial infarction. Science 2007;316:1491-3.

38. Visel A, Zhu Y, May D, et al. Targeted deletion of the 9p21 non-coding coronary artery disease risk interval in mice. Nature 2010;464:409-12.

39. Anderson CD, Biffi A, Rost NS, Cortellini L, Furie KL, Rosand J. Chromosome 9p21 in ischemic stroke: population structure and meta-analysis. Stroke 2010;41:1123-31.

40. Helgadottir A, Thorleifsson G, Magnusson KP, et al. The same sequence variant on 9p21 associates with myocardial infarction, abdominal aortic aneurysm and intracranial aneurysm. Nat Genet 2008;40:217-24.

41. Ganesh SK, Zakai NA, van Rooij FJ, et al. Multiple loci influence erythrocyte phenotypes in the CHARGE Consortium. Nat Genet 2009;41:1191-8.

42. International Consortium for Blood Pressure Genome-Wide Association Studies. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. Nature 2011;478:103-9.

43. Teslovich TM, Musunuru K, Smith AV, et al. Biological, clinical and population relevance of 95 loci for blood lipids. Nature 2010;466:707-13.

44. Tobacco and Genetics Consortium. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. Nat Genet 2010;42:441-7.

45. Voight BF, Scott LJ, Steinthorsdottir V, et al. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. Nat Genet 2010;42:579-89.

46. Dupuis J, Langenberg C, Prokopenko I, et al. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. Nat Genet 2010;42:105-16.

47. Speliotes EK, Willer CJ, Berndt SI, et al. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nat Genet 2010;42:937-48.

48. Heard-Costa NL, Zillikens MC, Monda KL, et al. NRXN3 is a novel locus for waist circumference: a genome-wide association study from the CHARGE Consortium. PLoS Genet 2009;5(6):e1000539.

49. Dehghan A, Yang Q, Peters A, et al. Association of novel genetic loci with circulating fibrinogen levels: a genome-wide association study in 6 population-based cohorts. Circ Cardiovasc Genet 2009;2:125-33.

50. Elliott P, Chambers JC, Zhang W, et al. Genetic loci associated with C-reactive protein levels and risk of coronary heart disease. JAMA 2009;302:37-48.

51. Barbalic M, Dupuis J, Dehghan A, et al.

The New England Journal of Medicine
Downloaded from nejm.org at VIRGINIA COMMONWEALTH UNIV on December 1, 2011. For personal use only. No other uses without permission.
Copyright © 2011 Massachusetts Medical Society. All rights reserved.

Large-scale genomic studies reveal central role of ABO in sP-selectin and sICAM-1 levels. Hum Mol Genet 2010;19:1863-72.

52. Mälarstig A, Buil A, Souto JC, et al. Identification of ZNF366 and PTPRD as novel determinants of plasma homocysteine in a family-based genome-wide association study. Blood 2009;114:1417-22.

53. Bis JC, Kavousi M, Franceschini N, et al. Meta-analysis of genome-wide association studies from the CHARGE consortium identifies common variants associated with carotid intima media thickness and plaque. Nat Genet 2011;43:940-7.

54. Linsel-Nitschke P, Götz A, Erdmann J, et al. Lifelong reduction of LDL-cholesterol related to a common variant in the LDL-receptor gene decreases the risk of coronary artery disease — a mendelian randomisation study. PLoS ONE 2008; 3(8):e2986.

55. Kamstrup PR, Tybjaerg-Hansen A, Steffensen R, Nordestgaard BG. Genetically elevated lipoprotein(a) and increased risk of myocardial infarction. JAMA 2009;301:2331-9.

56. C Reactive Protein Coronary Heart Disease Genetics Collaboration (CCGC). Association between C reactive protein and coronary heart disease: mendelian randomisation analysis based on individual participant data. BMJ 2011;342:d548.

57. Ji W, Foo JN, O'Roak BJ, et al. Rare independent mutations in renal salt handling genes contribute to blood pressure variation. Nat Genet 2008;40:592-9.

58. Romeo S, Pennacchio LA, Fu Y, et al. Population-based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase HDL. Nat Genet 2007;39:513-6.

59. Cohen JC, Boerwinkle E, Mosley TH Jr, Hobbs HH. Sequence variations in *PCSK9*, low LDL, and protection against coronary heart disease. N Engl J Med 2006; 354:1264-72.

60. Harismendy O, Bansal V, Bhatia G, et al. Population sequencing of two endocannabinoid metabolic genes identifies rare and common regulatory variants associated with extreme obesity and metabolite level. Genome Biol 2010;11:R118.

61. Johansen CT, Wang J, Lanktree MB, et al. Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia. Nat Genet 2010;42: 684-7.

62. Choi M, Scholl UI, Ji W, et al. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proc Natl Acad Sci U S A 2009;106:19096-101.

63. Musunuru K, Pirruccello JP, Do R, et al. Exome sequencing, ANGPTL3 mutations, and familial combined hypolipidemia. N Engl J Med 2010;363:2220-7.

64. Norton N, Li D, Rieder MJ, et al. Genome-wide studies of copy number variation and exome sequencing identify rare variants in BAG3 as a cause of dilated cardiomyopathy. Am J Hum Genet 2011;88: 273-82.

65. Roach JC, Glusman G, Smit AF, et al. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science 2010;328:636-9.

66. Rios J, Stein E, Shendure J, Hobbs HH, Cohen JC. Identification by whole-genome resequencing of gene defect responsible for severe hypercholesterolemia. Hum Mol Genet 2010;19:4313-8.

67. Harismendy O, Notani D, Song X, et al. 9p21 DNA variants associated with coronary artery disease impair interferon-γ signaling response. Nature 2011;470:264-8.

68. Musunuru K, Strong A, Frank-Kamenetsky M, et al. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. Nature 2010;466:714-9.

69. Cohen J, Pertsemlidis A, Kotowski IK, Graham R, Garcia CK, Hobbs HH. Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in PCSK9. Nat Genet 2005;37: 161-5.

70. Itzhaki I, Maizels L, Huber I, et al. Modelling the long QT syndrome with induced pluripotent stem cells. Nature 2011;471:225-9.

71. Yazawa M, Hsueh B, Jia X, et al. Using induced pluripotent stem cells to investigate cardiac phenotypes in Timothy syndrome. Nature 2011;471:230-4.

72. Palomaki GE, Melillo S, Neveux L, et al. Use of genomic profiling to assess risk for cardiovascular disease and identify individualized prevention strategies — a targeted evidence-based review. Genet Med 2010;12:772-84.

73. Ripatti S, Tikkanen E, Orho-Melander M, et al. A multilocus genetic risk score for coronary heart disease: case-control and prospective cohort analyses. Lancet 2010;376:1393-400.

74. Pepe MS, Gu JW, Morris DE. The potential of genes and other markers to inform about risk. Cancer Epidemiol Biomarkers Prev 2010;19:655-65.

75. Bell CJ, Dinwiddie DL, Miller NA, et al. Carrier testing for severe childhood recessive diseases by next-generation sequencing. Sci Transl Med 2011;3:65ra4.

76. Ashley EA, Butte AJ, Wheeler MT, et al. Clinical assessment incorporating a personal genome. Lancet 2010;375:1525-35.

77. Wang L, McLeod HL, Weinshilboum RM. Genomics and drug response. N Engl J Med 2011;364:1144-53.

78. Liggett SB, Mialet-Perez J, Thaneemit-Chen S, et al. A polymorphism within a conserved beta(1)-adrenergic receptor motif alters cardiac function and beta-blocker response in human heart failure. Proc Natl Acad Sci U S A 2006;103:11288-93.

79. The SEARCH Collaborative Group. *SLCO1B1* variants and statin-induced myopathy — a genomewide study. N Engl J Med 2008;359:789-99.

80. International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. Nature 2010;467:52-8.

Copyright © 2011 Massachusetts Medical Society.

RECEIVE THE *JOURNAL*'S TABLE OF CONTENTS EACH WEEK BY E-MAIL

To receive the table of contents of the *Journal* by e-mail every Wednesday evening, sign up at **NEJM.org.**

The New England Journal of Medicine
Downloaded from nejm.org at VIRGINIA COMMONWEALTH UNIV on December 1, 2011. For personal use only. No other uses without permission.
Copyright © 2011 Massachusetts Medical Society. All rights reserved.

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/24260532

Genomics of Lung Cancer

Article *in* Proceedings of the American Thoracic Society · May 2009

DOI: 10.1513/pats.200807-076LC · Source: PubMed

CITATIONS
40

READS
112

3 authors, including:

Alain charles Borczuk
Northwell Health

361 PUBLICATIONS **22,684** CITATIONS

SEE PROFILE

Rebecca Toonkel
Florida International University

53 PUBLICATIONS **978** CITATIONS

SEE PROFILE

All content following this page was uploaded by Alain charles Borczuk on 26 June 2014.

The user has requested enhancement of the downloaded file.

Genomics of Lung Cancer

Alain C. Borczuk[1], **Rebecca L. Toonkel**[2], **and Charles A. Powell**[2]

[1]Department of Pathology, and [2]Division of Pulmonary and Critical Care Medicine, Columbia University College of Physicians and Surgeons, New York, New York

This review focuses on recent research using genomics to examine lung carcinogenesis, histologic differentiation, and progression.

Keywords: adenocarcinoma; invasion; genomics; microarray; development

Lung cancer is the leading cause of cancer death in both men and women in the United States, despite its incidence being less than that of prostate cancer in men and breast cancer in women. With 166,000 deaths expected in 2008, the sum total of lung cancer deaths exceeds those of prostate, breast, and colon cancer combined (1). Prostate, breast, and colorectal carcinoma have all demonstrated significant improvements in 5-year survival over time and are currently 99, 88, and 64%, respectively. In contrast, the 5-year survival rate for lung cancer has remained relatively stable at 15%.

There are several potential explanations for the disparity between lung cancer survival and that of other common tumors, including late detection and histologic heterogeneity. Currently, over 75% of new lung cancer diagnoses are made in patients presenting with distant or regional metastatic disease (2). Histologically, prostate, breast, and colorectal carcinomas are uniformly adenocarcinoma and treatment is primarily determined by clinical stage, at times modified by results of molecular assays (3, 4). In contrast, only 30% of lung carcinoma is adenocarcinoma. Yet for the most part, until recently, non–small cell lung cancers (squamous, large cell, adenocarcinoma) were treated similarly, regardless of the biological heterogeneity associated with histology. It is likely that poor historical lung cancer response rates may in part be attributable to a relatively homogenous approach to a heterogeneous disease. Ongoing efforts are underway to identify clinically relevant biological properties of tumors that will facilitate individualized lung cancer treatment.

This research takes advantage of technical advances that allow rapid high-throughput assays to interrogate the genome (mRNA, microRNA, copy number, mutation analyses), proteome, and epigenome. In this review, we will focus on transcriptional profiling studies that have advanced our understanding of malignant transformation of lung epithelial cells and of lung cancer differentiation and progression. These advances have direct relevance to the goals of early detection, accurate prognostic assessment, and targeted therapy.

FIELD CARCINOGENESIS

Fifteen percent of lifetime smokers develop lung cancer, but 10% of lung cancers occur in never-smokers (5). In nonsmokers, exposure to secondhand smoke or to other lung carcinogens such as radon, asbestos, arsenic, or air pollution may be contributory. In both smokers and nonsmokers, genetic polymorphisms in genes associated with carcinogen metabolism, DNA damage repair, and cell cycle control may influence lung cancer susceptibility and modify the injury response associated with exposure to the dozens of carcinogens contained in tobacco smoke (6, 7). While the heritable component of lung cancer risk due to these genetic polymorphisms is most notable in early-onset cases occurring with less total smoke exposure, due to the limited number of genes studied, the overall attributable risk remains small even in this group of patients (7, 8). Recently, three genome-wide association studies identified a locus on chromosome 15q associated with lung cancer (9–11). The common locus variant near rs1051730 contains three genes encoding subunits of the nicotinic acetylcholine receptor (nAChR), CHRNA3, CHRNA5, and CHRNB4. The nAChR holoreceptor is a pentamer surrounding a gated channel responsive to ligands such as acetylcholine, nicotine, or its highly carcinogenic derivative 4(methylnitrosamine)-1-(3-pyridyl)-1-butanone (12). It is unclear whether the gene variant mediates susceptibility predominantly through influencing smoking behavior and nicotine dependence or through direct effects on lung tumor growth and suppression of apoptosis (13).

Because of tobacco smoke exposure, the bronchial epithelium of smokers is subject to field cancerization with alterations of airway cell DNA. Such alterations bring about oncogene activation, tumor suppressor gene silencing, and widespread loss of heterozygosity, all of which drive distinct gene expression signatures. While it is unclear which molecular alterations are required for neoplasia, in some instances, field carcinogenesis leads to malignant transformation of lung cancer progenitor cells. This malignant transformation may also require contributions from other systemic cell types such as bone marrow–derived stem cells (14, 15).

Microarray expression profiling has provided important information about cigarette smoke–associated lung carcinogenesis. Recently, we and others used gene expression profiling to identify gene signatures associated with cigarette smoking and lung adenocarcinoma in smokers and nonsmokers (16–20). Together, these results suggest that distinct cell transformation pathways and stromal tissue responses are involved in adenocarcinoma arising in smokers and nonsmokers (21). Observational studies using microarray analyses to profile bronchial epithelial cells obtained from bronchoscopic brushings of healthy smokers and nonsmokers identified overexpression of antioxidant genes and genes associated with xenobiotic metabolism by cells obtained from smokers (22, 23). However, these changes appeared to be potentially reversible when evaluated in a comparison of current versus former smokers. This transient genomic response to injury has been confirmed by *in vitro* studies using human bronchial epithelial cells exposed to cigarette smoke condensate and by *in vivo* rodent models of cigarette smoke exposure (24).

In addition, a field defect in smokers closely associated with lung cancer was reported by Spira and colleagues, who acquired gene expression signatures from nonmalignant bronchial cell brushings in a large number of smokers with and without lung

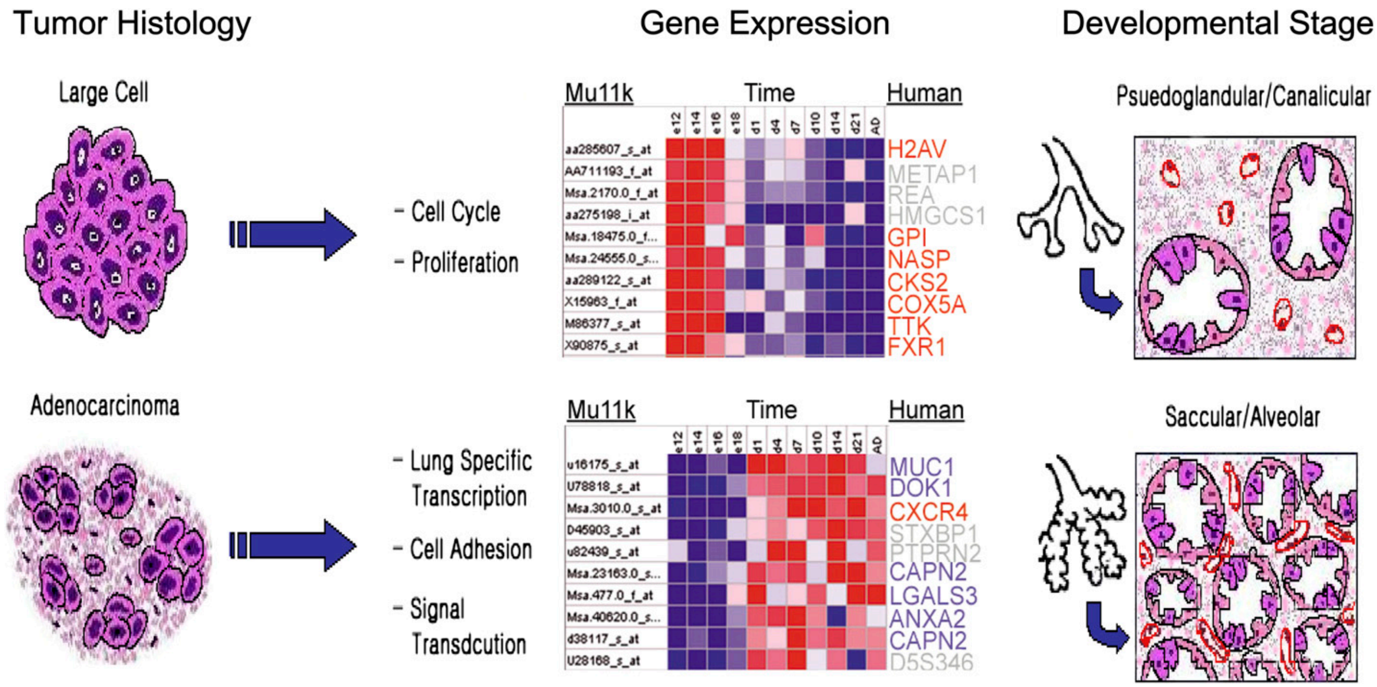
---

(*Received in original form July 31, 2008; accepted in final form November 6, 2008*)

Supported by the NIH (1RO1CA120174 to C.A.P.), the American Cancer Society (RSG-CNE-108857), Joan's Legacy Foundation, and by the Flight Attendants Medical Research Institute.

Correspondence and requests for reprints should be addressed to Charles A. Powell, M.D., Division of Pulmonary and Critical Care Medicine, Columbia University Medical Center, 630 West 168th Street, Box 91, New York, NY, 10032. E-mail: cap6@columbia.edu

Proc Am Thorac Soc Vol 6. pp 152–158, 2009
DOI: 10.1513/pats.200807-076LC
Internet address: www.atsjournals.org


Tumor Histology
Gene Expression
Developmental Stage
Large Cell
- Cell Cycle
- Proliferation
Mu11k
Time
Human
e12
e14
e16
e18
d1
d4
d7
d10
d14
d21
AD
aa285607_s_at
AA711193_f_at
Msa.2170.0_f_at
aa275198_i_at
Msa.18475.0_f...
Msa.24555.0_s...
aa289122_s_at
X15963_f_at
M86377_s_at
X90875_s_at
H2AV
METAP1
REA
HMGCS1
GPI
NASP
CKS2
COX5A
TTK
FXR1
Psuedoglandular/Canalicular
Adenocarcinoma
- Lung Specific Transcription
- Cell Adhesion
- Signal Transdcution
Mu11k
Time
Human
e12
e14
e16
e18
d1
d4
d7
d10
d14
d21
AD
u16175_s_at
U78818_s_at
Msa.3010.0_s_at
D45903_s_at
u82439_s_at
Msa.23163.0_s...
Msa.477.0_f_at
Msa.40620.0_s...
d38117_s_at
U28168_s_at
MUC1
DOK1
CXCR4
STXBP1
PTPRN2
CAPN2
LGALS3
ANXA2
CAPN2
D5S346
Saccular/Alveolar

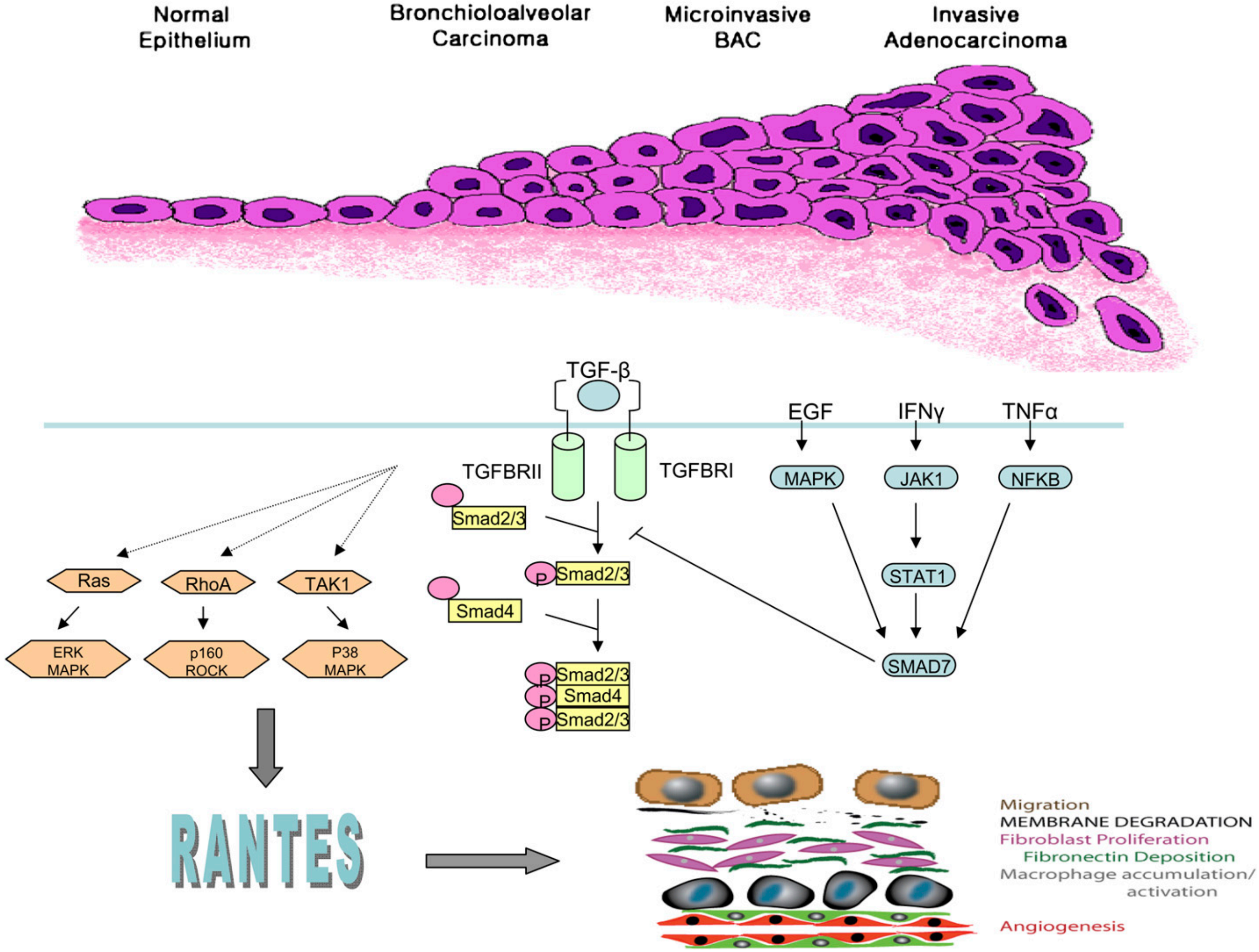

Normal Epithelium
Bronchioloalveolar Carcinoma
Microinvasive BAC
Invasive Adenocarcinoma
TGF-β
EGF
IFNγ
TNFα
TGFBRII
TGFBRI
MAPK
JAK1
NFKB
Smad2/3
P Smad2/3
Smad4
STAT1
SMAD7
Ras
RhoA
TAK1
ERK MAPK
p160 ROCK
P38 MAPK
P Smad2/3
P Smad4
P Smad2/3
RANTES
Migration
MEMBRANE DEGRADATION
Fibroblast Proliferation
Fibronectin Deposition
Macrophage accumulation/ activation
Angiogenesis

Figure 1. Comparison of gene expression of human lung tumors with genes temporally expressed during lung development suggests a recapitulation of developmentally regulated pathways. Nearest neighbor analysis was used to classify and rank murine (Mu11k) gene markers associated with two groups associated with lung development stage. Each *row* in the *heat map* (*middle panel*) shows gene expression over time for murine genes. Lung development stages represented by Group 1 (Days E12–E18, *top panel*) and Group 2 (Postnatal Day 1–Adult, *lower panel*) highly expressed genes are pseudoglandular/canalicular and terminal sac/alveolar, respectively. The corresponding human histology gene orthologs for large cell carcinoma are indicated in *red*, adenocarcinoma in *blue*. Specifically, genes expressed by large cell carcinoma mirror those expressed during the early pseudoglandular and canalicular stages of lung development, while those expressed by adenocarcinoma are predominantly expressed during the later terminal sac and alveolar stages.

cancer and observed that signatures from uninvolved cells in patients with cancer were distinct from those obtained from healthy smokers (25). This suggests that an epithelial "field defect" typically associated with cigarette smoke exposure may be specific for cancer. If validated prospectively, molecular testing of epithelial cells acquired throughout the aerodigestive tract may provide useful information for risk assessment.

GENOMICS OF LUNG CANCER DIFFERENTIATION AND PROGRESSION

Lung Developmental Pathways in Lung Cancer

Current paradigms suggest that lung carcinomas arise from pluripotent stem and progenitor cells capable of differentiation into one or several histologic cell types. These paradigms suggest that lung tumor cell ontology is determined by the consequences of gene transcriptional activation and/or repression events that recapitulate embryonic lung development (26, 27). The hypothesis that lung cancer arises from aberrant expression of genes involved in lung development is supported by gene expression studies demonstrating similarities between signatures obtained from human lung tumors and signatures characteristic of normal lung development (Figure 1). In an analysis of 32 non–small cell lung carcinoma (NSCLC) specimens and 7 normal specimens, unsupervised hierarchical analysis segregated tumors on the basis of histologic type and differentiation (28). Supervised clustering analysis of tumors identified numerous genes with known important function in embryonic lung development. Comparison of human lung tumor histology classifiers with genes temporally activated during mouse lung development reveals that genes expressed by large cell carcinoma (LCC) are similarly expressed during the early pseudoglandular and canalicular stages of lung development, while those expressed by adenocarcinoma mirror those expressed during the later terminal sac and alveolar stages. In addition to highlighting the expression of proliferation-associated genes by LCC and of differentiation-associated genes by adenocarcinoma, these results suggest a recapitulation of developmentally regulated pathways in lung tumors.

Similar observations were reported by Bonner and colleagues, who compared gene expression profiles of human lung tumors with those obtained from a murine model of lung adenocarcinoma (29). Liu and colleagues (30) found that genes associated with early lung development were more often expressed by small cell lung carcinoma (SCLC) and by tumors associated with poor prognosis. In addition, Glinsky and colleagues reported that a gene signature of "stemness" derived from BMI-1–regulated genes in normal stem cells is associated with metastasis and survival in several tumor types, including NSCLC (31). Taken together, these observations suggest that poor differentiation is linked to molecular parameters of early development representing lung stem and progenitor cell programs, and that gene signatures of these phenotypes are important for lung cancer differentiation, progression, and clinical outcome.

Neuroendocrine Tumors

One of the first studies using gene expression profiling to examine lung cancer histologic diversity compared carcinoids and small cell carcinoma tumors (32). Despite sharing neuroendocrine differentiation, the molecular profiles suggested different cells of origin for these tumor types. This molecular distinction was supported by a subsequent study from Bhattacharjee and colleagues (33) showing two distinct clusters of small cell carcinoma and carcinoid tumors, with small cell carcinoma characterized by frequent expression of proliferation markers such as MCM2, PCNA, MCM6, and thymidylate synthase. This was later confirmed by Virtanen and coworkers (34), who identified a carcinoid subgroup distinct from high-grade neuroendocrine carcinomas. Interestingly, large cell neuroendocrine carcinoma, a high-grade neuroendocrine carcinoma, was not readily distinguishable from small cell carcinoma by gene expression profiling. This observation suggested that despite being classified as an NSCLC subtype, large cell neuroendocrine carcinoma bears more similarity to small cell carcinoma than to other NSCLC subtypesn and illustrates how gene expression profiling can explain similar clinical features observed in histologically distinct tumors.

Squamous Cell Carcinoma

Researchers have examined gene expression profiles of squamous cell carcinomas (SCC) of the lung to address three major areas of concern: (*1*) to account for the histologic heterogeneity of NSCLC and to differentiate squamous cell carcinoma from other subtypes, (*2*) to identify prognostic factors specific to squamous cell carcinoma, and (*3*) to aid in distinguishing primary pulmonary SCC from metastatic SCC of the head and neck (HNSCC).

Several studies have identified gene expression profiles that distinguish pulmonary adenocarcinoma from pulmonary SCC (35–38). Notably, a shared set of genes encoding cytokeratins and cell adhesion proteins consistently distinguish squamous cell carcinoma from adenocarcinoma. These classifiers include cytokeratin 14 and components of desmosomes, hemi-desmosomes, and gap junctions, such as integrins ITGB4, desmocollin, and desmoplakin.

A consistent finding of unsupervised clustering analysis in pulmonary SCC is the generation of two clusters not attributable to histology, differentiation, or stage (39, 40). Rather, these clusters were enriched in prognosis associated genes encoding

Figure 2. Classification of lung adenocarcinoma reflects progression along a spectrum of invasiveness spanning epithelial atypia, bronchioloalveolar adenocarcinoma (BAC), BAC with foci of microinvasion, and invasive adenocarcinoma. In cells with repression of TGFβ-RII (*lower panel*), signaling through the Smad pathway is abrogated (*right*). Invasiveness is mediated through noncanonical signaling pathways that require RANTES (*left*).

proteins with function in the cell cycle and proliferation. Larson and colleagues (41) compared gene expression profiles of node-negative SCC tumors (N0) with signatures obtained from tumors with ipsalateral hilar or peribronchial nodal involvement (N1) via direct nodal invasion and with signatures obtained from N1 tumors with regional lymph node metastases via lymphatics, and found that N1 tumors with direct lymphatic invasion were similar to N0 tumors and distinct from N1 tumors with lymphatic spread, both in terms of molecular features and prognosis.

Although immunohistochemistry panels are effective in determining a site of primary origin for adenocarcinoma, equally effective panels do not exist for squamous cell carcinoma. Because primary SCC of the lung must be distinguished from metastatic HNSCC, this poses a diagnostic dilemma in patients with a history of HNSCC who subsequently develop a pulmonary nodule. Toward this end, Talbot and colleagues (42) developed a 500-gene classifier that distinguished pulmonary SCC from SCC of the tongue. In addition, Vachani and colleagues (43) used a training set of 18 HNSCC and 10 pulmonary squamous cell carcinomas to develop a 10-gene profile to distinguish these two tumor types. This classifier was validated on 122 subjects from previously published independent datasets, with 96% accuracy. Using real-time PCR, they correctly classified 12 independent samples as either of head and neck or pulmonary origin. These studies did not use microdissection; thus, it is possible that the signal from the host tissue may confound interpretation of gene signatures of tumor origin. However, if validated by others using microdissected specimens, these genomic classifiers hold promise for addressing important clinical staging issues in squamous cell carcinoma.

Adenocarcinoma

Somatic DNA alterations including mutations, amplifications, deletions, and translocations are common in tumors and are required for the activation of oncogenes and the inactivation of tumor suppressor genes that drive carcinogenesis. Large-scale efforts to systematically characterize these alterations in lung cancer and other tumors (44, 45) have provided insights into DNA structural alterations in non–small cell lung cancer. Adenocarcinoma, the most frequent histologic subtype of NSCLC, is characterized by multiple somatic DNA alterations. Recently, Weir and colleagues comprehensively examined DNA copy number alterations in 371 lung adenocarcinoma specimens (46) and found that the top focal regions of amplification and deletion included 14q13.3, 12q15, 8q24.21, 7p11.2, and 8q21.13. These results confirmed amplifications and deletions reported in other smaller studies, and also identified novel alterations such as amplification of the transcription factor TTF-1 on chromosome 14q13.3. *TTF-1* encodes thyroid transcription factor 1, a member of the Nk-2 homeobox family that binds to and activates the promoter of thyroid- and lung-specific genes. Interestingly, other groups using similar approaches also identified TTF-1 amplification as a frequent lung adenocarcinoma alteration, thus providing independent validation of this finding (47, 48). Together, these observations link lung tumor differentiation states and histologic grade with epithelial cell developmental pathways (28, 49), and suggest that these DNA amplification events are important in mediating lung cancer initiation, differentiation, and progression.

Lung Adenocarcinoma Invasion

Lung adenocarcinoma, the most frequent histologic type of NSCLC, is heterogeneous. Histologically, subclassification of adenocarcinoma is based upon World Health Organization (WHO) criteria, determined predominantly by cell morphology and growth pattern, along a spectrum including noninvasive bronchioloalveolar carcinoma (BAC), adenocarcinoma with mixed subtypes (AC-Mixed), and pure invasive adenocarcinoma (IAC) (50). Lung cancer metastasis represents the final step of a complex sequence composed of invasion (loss of cell–cell adhesion, increased cell motility, and basement membrane degradation), vascular intravasation and extravasation, establishment of a metastatic niche, and angiogenesis. Recent research has focused on characterizing the molecular mechanisms and clinical implications of adenocarcinoma invasion, the initial step of metastasis.

Paralleling malignancies in other organs, such as breast and cervix, where tumors are defined as noninvasive (*in situ* carcinoma), microinvasive (microscopic invasion), or as invasive carcinomas, the extent of the invasive component seen in lung adenocarcinoma is associated with clinical outcomes. Notably, this is most relevant in the lung for adenocarcinoma, because the noninvasive component is generally minimal in squamous cell carcinoma and often absent in large cell carcinoma. The clinical importance of lung adenocarcinoma invasion is supported by several recent studies (51–56) indicating that the risk of death in noninvasive BAC tumors is significantly lower than that of pure invasive tumors and in tumors with less than 0.6 cm of fibrosis or linear invasion. These results suggest that the prognosis of BAC is favorable, and suggest a similarly favorable prognosis for the subset of AC-Mixed subtype tumors with limited invasion (<6 mm).

We (57) and others (58–60) have used microarray gene expression profiling of lung adenocarcinoma tumors to identify signatures associated with histologic subtype and invasion. The results of unsupervised analyses are striking in that in each instance, tumors segregate into three major clades composed predominantly of BAC, AC-Mixed subtype, and pure invasive tumors, providing biological plausibility for the notion that these adenocarcinoma subtypes are distinct tumors. Taken together with the clinical data indicating distinct prognoses for each subtype, these studies have motivated consideration of revising the WHO lung adenocarcinoma classification scheme to reinforce the designation of purely noninvasive tumors and to create a designation for minimally invasive tumors.

To identify the molecular pathways important in mediating the acquisition of invasion by lung adenocarcinoma, we performed supervised analysis of mRNA microarray data to identify genes differentially expressed in noninvasive BAC and in AC-mixed type tumors. Among the genes differentially expressed in the progression from BAC to invasive tumors was the type II transforming growth factor β receptor (*TGF*β*RII*), which was less highly expressed by AC-Mixed and solid invasive tumors compared with BAC. This finding, which suggested that *TGF*β*RII* repression was required for lung adenocarcinoma invasion, is supported by genetic models combining targeted deletion of *TGF*β*RII* with other oncogenic events such as Apc mutation in colon tumors and Kras mutations in pancreatic and oropharyngeal carcinomas (61–63). The phenotypes of these *TGF*β*RII*-deficient cancer models clearly demonstrate the importance of TGF-β signaling in tumor invasion, yet the downstream signaling mechanisms are undefined.

We used a tumor cell invasion system and microarray analysis to identify and characterize downstream mediators of TGF-β signaling important for lung adenocarcinoma invasion (57). Among potential mediators identified was the chemokine CCL5 (RANTES [regulated on activation, normal T cell expressed, and presumably secreted]), which is upregulated by invasive tumors and *TGF*β*RII* knockdown cells. RANTES is involved in immunoregulatory and inflammatory processes and is transcribed and secreted not only by T cells, other inflammatory cells, and stromal cells, but also by tumor

TABLE 1. RECENT GENE EXPRESSION PREDICTORS OF NON–SMALL CELL LUNG CANCER OUTCOME

Classifier	Histology	Genes	Outcome	Platform
Duke Metagene (69)	NSCLC, Stages I, II, III	79	Recurrence	Affymetrix U133A
NCI Cytokine (72)	Adenocarcinoma, Stages I, II, III, IV	15	Lymph node metastasis and survival	RT-PCR
Milan 10 gene (73)	Adenocarcinoma, Stage I	E2F1, E2F4, HOXB7, HSPG2, MCM6, NUDCD1, RRM2, SERPINB5, SF3B1, SCGB3A1	Overall Survival	RT-PCR
Taiwan 5 gene (70)	NSCLC Stages I, II, III	*DUSP6, ERBB3, MMD, STAT1, LCK*	Overall and Relapse Free Survival	RT-PCR
Toronto 3 gene (74)	NSCLC Stages I, II, III	*CCR7, HIF1A, STX1A*	Overall Survival	RT-PCR

Definition of abbreviation: NSCLC = non–small cell lung cancer.

cells and normal bronchial epithelium. RANTES is a ligand for chemokine receptors CCR1, CCR3, CCR4, and CCR5, which are expressed on epithelial cells, macrophages, lymphocytes, dendritic cells, and stromal cells (64–67). Inhibition of RANTES signaling was found to be associated with abrogation of tumor invasion, suggesting that RANTES is required for invasion in *TGFβRII* repressed lung adenocarcinoma cells (Figure 2). The clinical significance of this pathway is further supported by the finding that tumor expression of both RANTES and CCR5 by a large panel of lung adenocarcinoma is associated with patient survival. These studies illustrate how information gained from global expression profiling of tumors can be used to identify key pathways and genes mediating tumor growth, invasion, and metastasis.

Gene Expression Predictors of Lung Cancer Survival

Proof of concept that tumor mRNA profiling provides clinically significant information regarding patient outcome after resection has been established. The promise of this approach to segregate tumors and to direct patient treatment based on these signatures is being prospectively tested in two breast cancer clinical trials. For NSCLC, several predictors have been developed, which for the most part are based on methodologically sound approaches that include independent validation. As indicated in Table 1, the results of these studies are heterogeneous both in terms of the number of genes in the predictors and in the specific genes included in each signature. This heterogeneity is expected, given differences in study design, assay platform, tumor histology, and patient selection. Subsequently, Shedden and colleagues published the results of a large, multicenter, blinded evaluation of several genomic signatures of prognosis in 442 adenocarcinomas (68). The addition of clinical covariates enhanced the performance of novel signatures, which performed slightly better than previously published signatures (69, 70). None of the signatures performed significantly better than the others. This landmark study is a model for the meticulous handling of challenges inherent in translational cancer genomic studies, and for its vast data repository of clinical and pathologically annotated data that now is available for testing of new signatures and hypotheses. It remains unclear which genomic predictor of prognosis is best or whether specific genes or entire signatures are most important in predicting outcome. Independent prospective evaluation of the predictive accuracy of these signatures, prospective clinical trials, and application to small biopsy specimens (71) are appropriate and will be required to extend this important area of research.

Gene expression profiling studies of lung tumors not only provide important information regarding prognosis and survival, but also help to identify potential therapeutic targets by offering key insights into how genetic alterations affect lung tumorigensis and metastasis. As medicine becomes increasingly reliant on genomic information, these studies hold much promise for the diagnosis and treatment of lung cancer.

Conflict of Interest Statement: A.C.B. is a co-inventor involved in a patent application "Gene expression profiles correlating with histology and prognosis" measuring gene expression of a set of genes with positive correlation with histologic subtype and prognosis. To date, no financial benefit has been derived from this application. R.L.T. does not have a financial relationship with a commercial entity that has an interest in the subject of this manuscript. C.A.P. received $125,000 in 2006–2007 from Philip Research, NA as research grants and is co-inventor for a pending patent filed by Columbia University for expression profiles correlating with histology and prognosis.

References

1. Jemal A, Siegel R, Ward E, Hao Y, Xu J, Murray T, Thun MJ. Cancer statistics, 2008. *CA Cancer J Clin* 2008;58:71–96.
2. Jemal A, Siegel R, Ward E, Murray T, Xu J, Smigal C, Thun MJ. Cancer statistics, 2006. *CA Cancer J Clin* 2006;56:106–130.
3. Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, Baehner FL, Walker MG, Watson D, Park T, *et al.* A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N Engl J Med* 2004;351:2817–2826.
4. Romond EH, Perez EA, Bryant J, Suman VJ, Geyer CE Jr, Davidson NE, Tan-Chiu E, Martino S, Paik S, Kaufman PA, *et al.* Trastuzumab plus adjuvant chemotherapy for operable her2-positive breast cancer. *N Engl J Med* 2005;353:1673–1684.
5. Spitz MR, Wei Q, Dong Q, Amos CI, Wu X. Genetic susceptibility to lung cancer: The role of DNA damage and repair. *Cancer Epidemiol Biomarkers Prev* 2003;12:689–698.
6. Hecht SS. Cigarette smoking and lung cancer: chemical mechanisms and approaches to prevention. *Lancet Oncol* 2002;3:461–469.
7. Schwartz AG, Prysak GM, Bock CH, Cote ML. The molecular epidemiology of lung cancer. *Carcinogenesis* 2007;28:507–518.
8. Vineis P, Anttila S, Benhamou S, Spinola M, Hirvonen A, Kiyohara C, Garte SJ, Puntoni R, Rannug A, Strange RC, *et al.* Evidence of gene gene interactions in lung carcinogenesis in a large pooled analysis. *Carcinogenesis* 2007;28:1902–1905.
9. Amos CI, Wu X, Broderick P, Gorlov IP, Gu J, Eisen T, Dong Q, Zhang Q, Gu X, Vijayakrishnan J, *et al.* Genome-wide association scan of tag snps identifies a susceptibility locus for lung cancer at 15q25.1. *Nat Genet* 2008;40:616–622.
10. Hung RJ, McKay JD, Gaborieau V, Boffetta P, Hashibe M, Zaridze D, Mukeria A, Szeszenia-Dabrowska N, Lissowska J, Rudnai P, *et al.* A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. *Nature* 2008;452:633–637.
11. Thorgeirsson TE, Geller F, Sulem P, Rafnar T, Wiste A, Magnusson KP, Manolescu A, Thorleifsson G, Stefansson H, Ingason A, *et al.* A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. *Nature* 2008;452:638–642.
12. Schuller HM, Orloff M. Tobacco-specific carcinogenic nitrosamines: ligands for nicotinic acetylcholine receptors in human lung cancer cells. *Biochem Pharmacol* 1998;55:1377–1384.
13. Lam DC, Girard L, Ramirez R, Chau WS, Suen WS, Sheridan S, Tin VP, Chung LP, Wong MP, Shay JW, *et al.* Expression of nicotinic acetylcholine receptor subunit genes in non-small-cell lung cancer reveals differences between smokers and nonsmokers. *Cancer Res* 2007;67:4638–4647.
14. Houghton J, Stoicov C, Nomura S, Rogers AB, Carlson J, Li H, Cai X, Fox JG, Goldenring JR, Wang TC. Gastric cancer originating from bone marrow-derived cells. *Science* 2004;306:1568–1571.
15. Avital I, Moreira AL, Klimstra DS, Leversha M, Papadopoulos EB, Brennan M, Downey RJ. Donor-derived human bone marrow cells contribute to solid organ cancers developing after bone marrow transplantation. *Stem Cells* 2007;25:2903–2909.
16. Powell CA, Spira A, Derti A, DeLisi C, Liu G, Borczuk A, Busch S, Sahasrabudhe S, Chen Y, Sugarbaker D, *et al.* Gene expression in

lung adenocarcinomas of smokers and nonsmokers. *Am J Respir Cell Mol Biol* 2003;29:157–162.

17. Woenckhaus M, Klein-Hitpass L, Grepmeier U, Merk J, Pfeifer M, Wild P, Bettstetter M, Wuensch P, Blaszyk H, Hartmann A, *et al.* Smoking and cancer-related gene expression in bronchial epithelium and non-small-cell lung cancers. *J Pathol* 2006;210:192–204.
18. Landi MT, Dracheva T, Rotunno M, Figueroa JD, Liu H, Dasgupta A, Mann FE, Fukuoka J, Hames M, Bergen AW, *et al.* Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival. *PLoS One* 2008;3:e1651.
19. Miura K, Bowman ED, Simon R, Peng AC, Robles AI, Jones RT, Katagiri T, He P, Mizukami H, Charboneau L, *et al.* Laser capture microdissection and microarray expression analysis of lung adenocarcinoma reveals tobacco smoking- and prognosis-related molecular profiles. *Cancer Res* 2002;62:3244–3250.
20. Shigematsu H, Lin L, Takahashi T, Nomura M, Suzuki M, Wistuba II, Fong KM, Lee H, Toyooka S, Shimizu N, *et al.* Clinical and biological features associated with epidermal growth factor receptor gene mutations in lung cancers. *J Natl Cancer Inst* 2005;97:339–346.
21. Subramanian J, Govindan R. Lung cancer in never smokers: a review. *J Clin Oncol* 2007;25:561–570.
22. Hackett NR, Heguy A, Harvey BG, O'Connor TP, Luettich K, Flieder DB, Kaplan R, Crystal RG. Variability of antioxidant-related gene expression in the airway epithelium of cigarette smokers. *Am J Respir Cell Mol Biol* 2003;29:331–343.
23. Spira A, Beane J, Shah V, Liu G, Schembri F, Yang X, Palma J, Brody JS. Effects of cigarette smoke on the human airway epithelial cell transcriptome. *Proc Natl Acad Sci USA* 2004;101:10143–10148.
24. Gebel S, Gerstmayer B, Bosio A, Haussmann H-J, Miert EV, Muller T. Gene expression profiling in respiratory tissues from rats exposed to mainstream cigarette smoke. *Carcinogenesis* 2004;25:169–178.
25. Spira A, Beane JE, Shah V, Steiling K, Liu G, Schembri F, Gilman S, Dumas YM, Calner P, Sebastiani P, *et al.* Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer. *Nat Med* 2007;13:361–366.
26. Reya T, Morrison SJ, Clarke MF, Weissman IL. Stem cells, cancer, and cancer stem cells. *Nature* 2001;414:105–111.
27. Heguy A, O'Connor TP, Luettich K, Worgall S, Cieciuch A, Harvey B-G, Hackett NR, Crystal RG. Gene expression profiling of human alveolar macrophages of phenotypically normal smokers and nonsmokers reveals a previously unrecognized subset of genes modulated by cigarette smoking. *J Mol Med* 2006;84:318.
28. Borczuk AC, Gorenstein L, Walter KL, Assaad AA, Wang L, Powell CA. Non-small-cell lung cancer molecular signatures recapitulate lung developmental pathways. *Am J Pathol* 2003;163:1949–1960.
29. Bonner AE, Lemon WJ, Devereux TR, Lubet RA, You M. Molecular profiling of mouse lung tumors: association with tumor progression, lung development, and human lung adenocarcinomas. *Oncogene* 2004;23:1166–1176.
30. Liu H, Kho AT, Kohane IS, Sun Y. Predicting survival within the lung cancer histopathological hierarchy using a multi-scale genomic model of development. *PLoS Med* 2006;3:e232.
31. Glinsky GV, Berezovska O, Glinskii AB. Microarray analysis identifies a death-from-cancer signature predicting therapy failure in patients with multiple types of cancer. *J Clin Invest* 2005;115:1503–1521.
32. Anbazhagan R, Tihan T, Bornman DM, Johnston JC, Saltz JH, Weigering A, Piantadosi S, Gabrielson E. Classification of small cell lung cancer and pulmonary carcinoid by gene expression profiles. *Cancer Res* 1999;59:5119–5122.
33. Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, Ladd C, Beheshti J, Bueno R, Gillette M, *et al.* Classification of human lung carcinomas by mrna expression profiling reveals distinct adenocarcinoma subclasses. *Proc Natl Acad Sci USA* 2001;98:13790–13795.
34. Virtanen C, Ishikawa Y, Honjoh D, Kimura M, Shimane M, Miyoshi T, Nomura H, Jones MH. Integrated classification of lung tumors and cell lines by expression profiling. *Proc Natl Acad Sci USA* 2002;99: 12357–12362.
35. Boelens MC, van den Berg A, Vogelzang I, Wesseling J, Postma DS, Timens W, Groen HJ. Differential expression and distribution of epithelial adhesion molecules in non-small cell lung cancer and normal bronchus. *J Clin Pathol* 2007;60:608–614.
36. Kettunen E, Anttila S, Seppanen JK, Karjalainen A, Edgren H, Lindstrom I, Salovaara R, Nissen AM, Salo J, Mattson K, *et al.* Differentially expressed genes in nonsmall cell lung cancer: expression profiling of cancer-related genes in squamous cell lung cancer. *Cancer Genet Cytogenet* 2004;149:98–106.
37. Kuner R, Muley T, Meister M, Ruschhaupt M, Buness A, Xu EC, Schnabel P, Warth A, Poustka A, Sultmann H, *et al.* Global gene expression analysis reveals specific patterns of cell junctions in non-small cell lung cancer subtypes. *Lung Cancer* 2009;63:32–38.
38. Wang T, Hopkins D, Schmidt C, Silva S, Houghton R, Takita H, Repasky E, Reed SG. Identification of genes differentially over-expressed in lung squamous cell carcinoma using combination of cDNA subtraction and microarray analysis. *Oncogene* 2000;19:1519–1528.
39. Inamura K, Fujiwara T, Hoshida Y, Isagawa T, Jones MH, Virtanen C, Shimane M, Satoh Y, Okumura S, Nakagawa K, *et al.* Two subclasses of lung squamous cell carcinoma with different gene expression profiles and prognosis identified by hierarchical clustering and non-negative matrix factorization. *Oncogene* 2005;24:7105–7113.
40. Larsen JE, Pavey SJ, Passmore LH, Bowman R, Clarke BE, Hayward NK, Fong KM. Expression profiling defines a recurrence signature in lung squamous cell carcinoma. *Carcinogenesis* 2007;28:760–766.
41. Larsen JE, Pavey SJ, Bowman R, Yang IA, Clarke BE, Colosimo ML, Hayward NK, Fong KM. Gene expression of lung squamous cell carcinoma reflects mode of lymph node involvement. *Eur Respir J* 2007;30:21–25.
42. Talbot SG, Estilo C, Maghami E, Sarkaria IS, Pham DK, Oc P, Socci ND, Ngai I, Carlson D, Ghossein R, *et al.* Gene expression profiling allows distinction between primary and metastatic squamous cell carcinomas in the lung. *Cancer Res* 2005;65:3063–3071.
43. Vachani A, Nebozhyn M, Singhal S, Alila L, Wakeam E, Muschel R, Powell CA, Gaffney P, Singh B, Brose MS, *et al.* A 10-gene classifier for distinguishing head and neck squamous cell carcinoma and lung squamous cell carcinoma. *Clin Cancer Res* 2007;13:2905–2915.
44. Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, Davies H, Teague J, Butler A, Stevens C, *et al.* Patterns of somatic mutation in human cancer genomes. *Nature* 2007;446:153–158.
45. Collins FS, Barker AD. Mapping the cancer genome: pinpointing the genes involved in cancer will help chart a new course across the complex landscape of human malignancies. *Sci Am* 2007;296: 50–57.
46. Weir BA, Woo MS, Getz G, Perner S, Ding L, Beroukhim R, Lin WM, Province MA, Kraja A, Johnson LA, *et al.* Characterizing the cancer genome in lung adenocarcinoma. *Nature* 2007;450:893–898.
47. Kendall J, Liu Q, Bakleh A, Krasnitz A, Nguyen KC, Lakshmi B, Gerald WL, Powers S, Mu D. Oncogenic cooperation and coamplification of developmental transcription factor genes in lung cancer. *Proc Natl Acad Sci USA* 2007;104:16663–16668.
48. Tanaka H, Yanagisawa K, Shinjo K, Taguchi A, Maeno K, Tomida S, Shimada Y, Osada H, Kosaka T, Matsubara H, *et al.* Lineage-specific dependency of lung adenocarcinomas on the lung development regulator ttf-1. *Cancer Res* 2007;67:6007–6011.
49. Borczuk AC, Powell CA. Expression profiling and lung cancer development. *Proc Am Thorac Soc* 2007;4:127–132.
50. Brambilla E, Travis WD, Colby TV, Corrin B, Shimosato Y. The new world health organization classification of lung tumours. *Eur Respir J* 2001;18:1059–1068.
51. Maeshima AM, Niki T, Maeshima A, Yamada T, Kondo H, Matsuno Y. Modified scar grade: a prognostic indicator in small peripheral lung adenocarcinoma. *Cancer* 2002;95:2546–2554.
52. Borczuk AC, Qian F, Eleazar J, Assad A, Kazeros A, Powell CA. Invasive size is an independent predictor of survival in pulmonary adenocarcinoma. *Am J Surg Pathol* (In press)
53. Rena O, Papalia E, Ruffini E, Casadio C, Filosso PL, Oliaro A, Maggi G. Stage I pure bronchioloalveolar carcinoma: recurrences, survival and comparison with adenocarcinoma of the lung. *Eur J Cardiothorac Surg* 2003;23:409–414.
54. Sakurai H, Maeshima A, Watanabe S, Suzuki K, Tsuchiya R, Maeshima AM, Matsuno Y, Asamura H. Grade of stromal invasion in small adenocarcinoma of the lung: histopathological minimal invasion and prognosis. *Am J Surg Pathol* 2004;28:198–206.
55. Suzuki K, Yokose T, Yoshida J, Nishimura M, Takahashi K, Nagai K, Nishiwaki Y. Prognostic significance of the size of central fibrosis in peripheral adenocarcinoma of the lung. *Ann Thorac Surg* 2000;69: 893–897.
56. Yim J, Zhu LC, Chiriboga L, Watson HN, Goldberg JD, Moreira AL. Histologic features are important prognostic indicators in early stages lung adenocarcinomas. *Mod Pathol* 2007;20:233–241.
57. Borczuk AC, Kim HK, Yegen HA, Friedman RA, Powell CA. Lung adenocarcinoma global profiling identifies type II transforming growth factor-beta receptor as a repressor of invasiveness. *Am J Respir Crit Care Med* 2005;172:729–737.

58. Takeuchi T, Tomida S, Yatabe Y, Kosaka T, Osada H, Yanagisawa K, Mitsudomi T, Takahashi T. Expression profile-defined classification of lung adenocarcinoma shows close relationship with underlying major genetic changes and clinicopathologic behaviors. *J Clin Oncol* 2006;24:1679–1688.
59. Motoi N, Szoke J, Riely GJ, Seshan VE, Kris MG, Rusch VW, Gerald WL, Travis WD. Lung adenocarcinoma: modification of the 2004 WHO mixed subtype to include the major histologic subtype suggests correlations between papillary and micropapillary adenocarcinoma subtypes, EGFR mutations and gene expression analysis. *Am J Surg Pathol* 2008;32:810–827.
60. Beer DG, Kardia SL, Huang CC, Giordano TJ, Levin AM, Misek DE, Lin L, Chen G, Gharib TG, Thomas DG, *et al.* Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat Med* 2002;8:816–824.
61. Ijichi H, Chytil A, Gorska AE, Aakre ME, Fujitani Y, Fujitani S, Wright CV, Moses HL. Aggressive pancreatic ductal adenocarcinoma in mice caused by pancreas-specific blockade of transforming growth factor-beta signaling in cooperation with active kras expression. *Genes Dev* 2006;20:3147–3160.
62. Lu SL, Herrington H, Reh D, Weber S, Bornstein S, Wang D, Li AG, Tang CF, Siddiqui Y, Nord J, *et al.* Loss of transforming growth factor-beta type II receptor promotes metastatic head-and-neck squamous cell carcinoma. *Genes Dev* 2006;20:1331–1342.
63. Munoz NM, Upton M, Rojas A, Washington MK, Lin L, Chytil A, Sozmen EG, Madison BB, Pozzi A, Moon RT, *et al.* Transforming growth factor beta receptor type ii inactivation induces the malignant transformation of intestinal neoplasms initiated by apc mutation. *Cancer Res* 2006;66:9837–9844.
64. Fraziano M, Cappelli G, Santucci M, Mariani F, Amicosante M, Casarini M, Giosue S, Bisetti A, Colizzi V. Expression of CCR5 is increased in human monocyte-derived macrophages and alveolar macrophages in the course of *in vivo* and *in vitro* mycobacterium tuberculosis infection. *AIDS Res Hum Retroviruses* 1999;15:869–874.
65. Kunkel EJ, Boisvert J, Murphy K, Vierra MA, Genovese MC, Wardlaw AJ, Greenberg HB, Hodge MR, Wu L, Butcher EC, *et al.* Expression of the chemokine receptors CCR4, CCR5, and CXCR3 by human tissue-infiltrating lymphocytes. *Am J Pathol* 2002;160:347–355.
66. Ma B, Liu W, Homer RJ, Lee PJ, Coyle AJ, Lora JM, Lee CG, Elias JA. Role of CCR5 in the pathogenesis of IL-13-induced inflammation and remodeling. *J Immunol* 2006;176:4968–4978.
67. van Deventer HW, O'Connor W Jr, Brickey WJ, Aris RM, Ting JP, Serody JS. C-C chemokine receptor 5 on stromal cells promotes pulmonary metastasis. *Cancer Res* 2005;65:3374–3379.
68. Shedden K, Taylor JM, Enkemann SA, Tsao MS, Yeatman TJ, Gerald WL, Eschrich S, Jurisica I, Giordano TJ, Misek DE, *et al.* Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. *Nat Med* 2008;14:822–827.
69. Potti A, Mukherjee S, Petersen R, Dressman HK, Bild A, Koontz J, Kratzke R, Watson MA, Kelley M, Ginsburg GS, *et al.* A genomic strategy to refine prognosis in early-stage non-small-cell lung cancer. *N Engl J Med* 2006;355:570–580.
70. Chen HY, Yu SL, Chen CH, Chang GC, Chen CY, Yuan A, Cheng CL, Wang CH, Terng HJ, Kao SF, *et al.* A five-gene signature and clinical outcome in non-small-cell lung cancer. *N Engl J Med* 2007;356: 11–20.
71. Borczuk AC, Shah L, Pearson GD, Walter KL, Wang L, Austin JH, Friedman RA, Powell CA. Molecular signatures in biopsy specimens of lung cancer. *Am J Respir Crit Care Med* 2004;170:167–174.
72. Seike M, Yanaihara N, Bowman ED, Zanetti KA, Budhu A, Kumamoto K, Mechanic LE, Matsumoto S, Yokota J, Shibata T, *et al.* Use of a cytokine gene expression signature in lung adenocarcinoma and the surrounding tissue as a prognostic classifier. *J Natl Cancer Inst* 2007; 99:1257–1269.
73. Bianchi F, Nuciforo P, Vecchi M, Bernard L, Tizzoni L, Marchetti A, Buttitta F, Felicioni L, Nicassio F, Di Fiore PP. Survival prediction of stage I lung adenocarcinomas by expression of 10 genes. *J Clin Invest* 2007;117:3436–3444.
74. Lau SK, Boutros PC, Pintilie M, Blackhall FH, Zhu C-Q, Strumpf D, Johnston MR, Darling G, Keshavjee S, Waddell TK, *et al.* Three-gene prognostic classifier for early-stage non small-cell lung cancer. *J Clin Oncol* 2007;25:5562–5569.

View publication stats

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/233770705

Abstract 2206: Genomics of Drug Sensitivity in Cancer (GDSC): A resource for therapeutic biomarker discovery in cancer cells.

Article *in* Nucleic Acids Research · November 2012

DOI: 10.1093/nar/gks1111 · Source: PubMed

CITATIONS
3,441

READS
1,728

17 authors, including:

Wanjuan Yang
Wellcome Sanger Institute
14 PUBLICATIONS **6,586** CITATIONS
SEE PROFILE

Jorge Soares
Pirum Systems
7 PUBLICATIONS **7,164** CITATIONS
SEE PROFILE

Patricia Greninger
Massachusetts General Hospital
86 PUBLICATIONS **13,796** CITATIONS
SEE PROFILE

Howard Lightfoot
Wellcome Sanger Institute
54 PUBLICATIONS **7,840** CITATIONS
SEE PROFILE

All content following this page was uploaded by Richard Thompson on 20 September 2014.

The user has requested enhancement of the downloaded file.

Published online 23 November 2012 *Nucleic Acids Research, 2013, Vol. 41, Database issue* **D955–D961**
doi:10.1093/nar/gks1111

Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells

Wanjuan Yang[1], Jorge Soares[1], Patricia Greninger[2], Elena J. Edelman[2], Howard Lightfoot[1], Simon Forbes[1], Nidhi Bindal[1], Dave Beare[1], James A. Smith[3], I. Richard Thompson[1], Sridhar Ramaswamy[2], P. Andrew Futreal[1], Daniel A. Haber[2,4], Michael R. Stratton[1], Cyril Benes[2], Ultan McDermott[1,]* and Mathew J. Garnett[1,]*

[1]Cancer Genome Project, Wellcome Trust Sanger Institute, Hinxton CB10 1SA, UK, [2]Center for Molecular Therapeutics, Massachusetts General Hospital Cancer Center, Harvard Medical School, Charlestown, MA 02129, USA, [3]Core Software Services, Wellcome Trust Sanger Institute, Hinxton CB10 1SA, UK and [4]Howard Hughes Medical Institute, Chevy Chase, MD 20815, USA

Received August 29, 2012; Revised October 15, 2012; Accepted October 20, 2012

ABSTRACT

Alterations in cancer genomes strongly influence clinical responses to treatment and in many instances are potent biomarkers for response to drugs. The Genomics of Drug Sensitivity in Cancer (GDSC) database (www.cancerRxgene.org) is the largest public resource for information on drug sensitivity in cancer cells and molecular markers of drug response. Data are freely available without restriction. GDSC currently contains drug sensitivity data for almost 75 000 experiments, describing response to 138 anticancer drugs across almost 700 cancer cell lines. To identify molecular markers of drug response, cell line drug sensitivity data are integrated with large genomic datasets obtained from the Catalogue of Somatic Mutations in Cancer database, including information on somatic mutations in cancer genes, gene amplification and deletion, tissue type and transcriptional data. Analysis of GDSC data is through a web portal focused on identifying molecular biomarkers of drug sensitivity based on queries of specific anticancer drugs or cancer genes. Graphical representations of the data are used throughout with links to related resources and all datasets are fully downloadable. GDSC provides a unique resource incorporating large drug sensitivity and genomic datasets to facilitate the discovery of new therapeutic biomarkers for cancer therapies.

INTRODUCTION

There is compelling evidence that alterations in cancer genomes can strongly influence clinical responses to anticancer therapies. Indeed, there are now several examples where genomic changes can be used as molecular biomarkers to identify patients most likely to benefit from a treatment. For example, the use of drugs to target the protein product of the *BCR–ABL* translocation in chronic myeloid leukemia, or the *BRAF* gene in malignant melanoma, has transformed the treatment of these diseases and substantially improved survival rates (1,2). Despite these notable successes, many cancer drugs in use or development have not been linked to specific genomic markers that could direct their clinical use to maximize patient benefit. Moreover, even among appropriately selected patients, a poorly explained range of clinical responses is observed (2,3). Thus, there exists a need for the development of new and improved biomarkers to guide therapies and ultimately improve clinical responses.

Recent years have seen significant advances in our understanding of the molecular nature of cancer (4). This has been driven in part by advances in high-throughput technologies and, in particular, DNA sequencing technologies that allow us to sequence on a scale that was previously unthinkable. In the near future, sequencing efforts will provide a complete description of the genomic changes that occur in many cancer subtypes. A complete list of the repertoire of cancer genes will provide profound insights into the origins, evolution and progression of cancer and will act as an impetus for the development of new cancer therapies.

---

*To whom correspondence should be addressed. Tel: +44 1223 494878; Fax: +44 1226 494919; Email: mg12@sanger.ac.uk
Correspondence may also be address to Ultan McDermott. Tel: +44 1223 494856; Fax: +44 1226 494919; Email: um1@sanger.ac.uk

© The Author(s) 2012. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by-nc/3.0/), which permits non-commercial reuse, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com.

To exploit this increased understanding, preclinical studies that link the genomic complexity of cancer with functional readouts such as drug sensitivity are required. Cancer cell lines derived from naturally occurring tumours have been generated from many different cancer types and in many respects recapitulate the tissue type and genomic context of cancer. They are a facile system for experimental manipulation and are a standard research tool in molecular biology and drug discovery. Significantly, several studies have used cancer cell lines to link pharmacological data with genomic information and helped define therapeutic biomarkers (5–7). Collectively, these studies have demonstrated that pharmacogenomic profiling in cancer cell lines can be used as a biomarker discovery platform to guide the development of new cancer therapies.

The Genomics of Drug Sensitivity in Cancer (GDSC) database (www.cancerRxgene.org) is designed to facilitate an increased understanding of the molecular features that influence drug response in cancer cells and which will enable the design of improved cancer therapies. GDSC holds and annotates large datasets on drug sensitivity in cancer cells and links these data to detailed genomic information to facilitate the discovery of molecular biomarkers of drug response. The website is designed to provide straightforward access to querying the database, and interactive graphical interfaces are used throughout to provide readily interpretable summaries of data and analyses.

DATABASE CONTENT

The GDSC database is based on three types of datasets as described in the following sections.

Cell line drug sensitivity data

Cancer cell line drug sensitivity data are generated from ongoing high-throughput screening performed by the Cancer Genome Project at the Wellcome Trust Sanger Institute (WTSI) and the Center for Molecular Therapeutics at Massachusetts General Hospital using a collection of >1000 cell lines (7). Compounds selected for screening are anticancer therapeutics encompassing both targeted agents and cytotoxic chemotherapeutics. They are comprised of approved drugs used in the clinic, drugs undergoing clinical development and in clinical trials and tool compounds in early phase development. They cover a wide range of targets and processes implicated in cancer biology including receptor tyrosine kinase signalling, cell cycle control, DNA damage response and the cytoskeleton. Compounds are sourced from commercial vendors or provided by collaborators in academia, biotech and the pharmaceutical industry.

Cell line drug sensitivity is measured using fluorescence-based cell viability assays following 72 h of drug treatment. Dose–response curves are fitted to fluorescence signal intensities over nine drug concentrations (2-fold dilution series) to derive a multi-parameter signature of drug response. Values reported on the website include the half maximal inhibitory concentration (IC_{50}), the slope of the dose–response curve and the area under the curve for each experiment.

The current release of GDSC (release 2, July 2012) includes drug sensitivity data for 138 anticancer compounds screened across a range of 329–668 cell lines per drug (mean = 525 cell lines per drug) representing 73 169 cell line–drug interactions. This is the largest public resource available on drug sensitivity in cancer cells. Screening is ongoing and the objective is to screen these compounds, as well as additional compounds in the future, across the entire collection of >1000 cell lines. Data release occurs every 4 months and with each release, these results are updated with new data for existing drugs, as well as data for newly screened drugs.

Genomic datasets for cell lines

The total collection available for screening includes >1000 different cancer cell lines. These have been selected to represent the spectrum of common and rare types of adult and childhood cancers of epithelial, mesenchymal and haematopoietic origin. The cell lines have been extensively genomically characterized as part of the cancer cell line project from the Cancer Genome Project at the WTSI. The genomic datasets currently available for each cell line include information on somatic mutations in 75 cancer genes, genome wide gene copy number for amplification and deletion, targeted screening for seven gene rearrangements, markers of microsatellite instability, tissue type and transcriptional data. Using various statistical approaches as described below, genomic datasets are used together with drug sensitivity data for each cell line to identify genomic biomarkers of drug response. Genomic datasets within GDSC are obtained and updated directly from the Catalogue of Somatic Mutations in Cancer (COSMIC) database, a comprehensive freely available resource for the annotation and presentation of somatic mutations in cancer (8).

Analysis of genomic features of drug sensitivity

An essential component of the GDSC database is the systematic integration of large-scale genomic and drug sensitivity datasets. To identify genomic markers of drug response, we currently use two complementary analytical approaches (7). A multivariate analysis of variance (MANOVA) is used to correlate drug sensitivity (IC_{50} values and slope of the dose–response curve) with genomic alterations in cancer including point mutations, amplifications and deletions of common cancer genes, cancer gene rearrangements and microsatellite instability. The MANOVA identifies individual genomic features associated with drug sensitivity and for each drug–gene association reports a size effect and statistical significance of the association.

We also apply elastic net regression, a penalized linear modelling technique, to identify multiple interacting genomic features influencing each drug response. Genomic data used in the elastic net analysis include all of those used in the MANOVA and also incorporate genome-wide transcriptional profiles and tissue type. The elastic net selects which of these features are associated with drug response as measured by IC_{50} values across the cell line panel. For each drug, a feature list is built comprised of mutations, transcripts and tissue with an effect size assigned to each.

A more detailed description of the different statistical analyses performed, as well as guidance on interpreting the results, can be found on the 'Help & Documentation' webpages under the 'statistical analysis' tab.

DATA ACCESS

Querying the GDSC database

The website is focused on presenting cell line drug sensitivity data and genomic correlates of drug sensitivity. Although data on the genomic characterization of the cell lines are available through the GDSC website, these data are presented in more detail within the COSMIC database.

To facilitate data interpretation, graphical representations with interactive features are used wherever possible. Querying the database is primarily based on either specific screening 'Compounds' or 'Cancer Genes' in the 'Browse our data' section of the homepage (Figure 1). Browsing by 'Compounds' displays a list of drug names together with their associated synonyms, putative therapeutic target(s), the number of cell lines screened for each drug (sample size) and date of the most recent data update for each compound. A link to the PUBCHEM database of chemical structures is provided (9). By clicking a specific drug name, users enter the individual drug page where drug sensitivity and genomic correlation data are presented.

Similarly, browsing 'Cancer Genes' leads to a list of cancer genes identified by their HUGO name. This page provides direct links to the COSMIC page for the gene and to the UniProt databases for further protein information (10). Clicking on the gene name accesses the drug sensitivity and genomic correlation data on the individual gene page.

It is also possible to query the database using a 'Search' function (Figure 1). The 'Search' box accepts queries based on compound (including synonyms), cancer gene or cell line name. An auto-completion feature enables users to quickly select their drug, gene or cell line of interest. The search result page lists matching compounds, cancer genes or cell lines with links to the detailed drug/gene page of the website. In the case of cell line matches, links are provided to detailed cell line information within COSMIC.

Data analysis and visualization

Screening data and genomic correlations are accessed through specific drug or gene pages (Figures 2a and 3). The top panel provides drug or gene information and links to PUBCHEM, COSMIC and UniProt databases as appropriate. Notably, the top panel also provides links to relevant help pages to explain the data and analyses performed. Additional information is also available from the 'Help & Documentation' link found in the header at the top of all pages. The actual screening data and analyses are presented in the bottom panel of a drug/gene page and are split into the following tabs: Volcano plot, Volcano data, Elastic net (drug pages only), Scatter plots and Download data.

A volcano plot is used to visualize the correlation of drug sensitivity data with genetic events as calculated using the MANOVA. The drug page shows a drug-specific volcano plot, which represents how different genomic changes influence response to a specific drug (Figure 2a). The gene page shows a gene-specific volcano plot, which represents the effect of a mutated cancer gene on the responses to all drugs analysed (Figure 3). For example, the

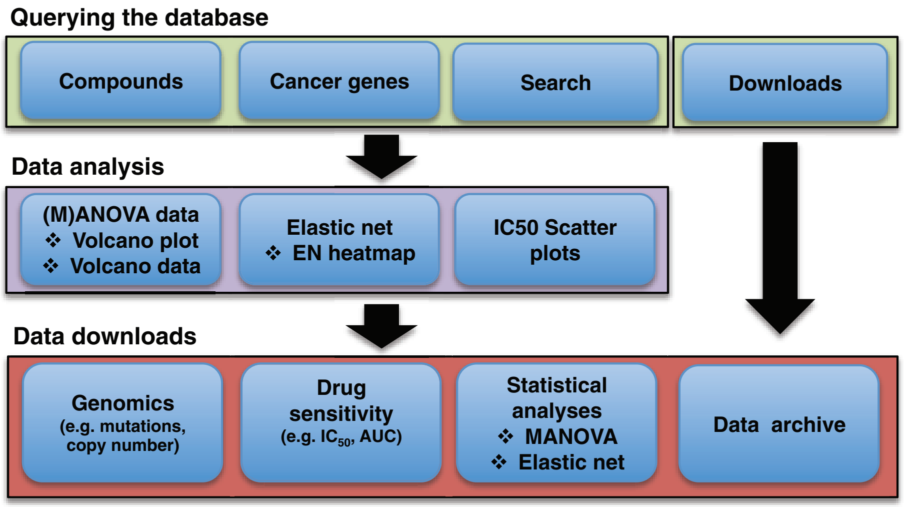
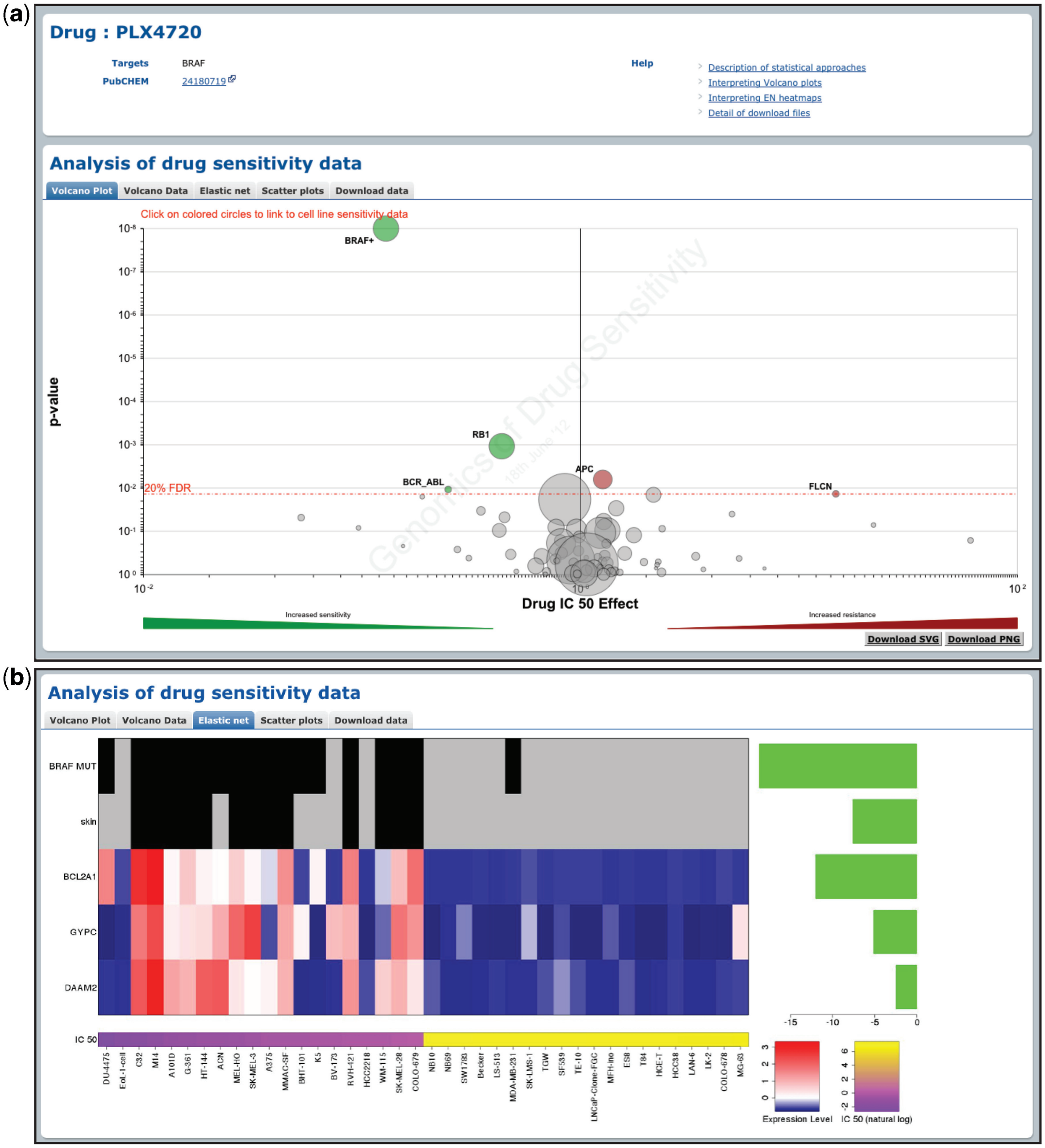


Figure 1. A schematic representation of the GDSC database structure and content. Data are accessed in a hierarchical fashion by either querying by screening compound or cancer gene of interest. This gives access to graphical representations of cell line drug sensitivity data and genomic correlations of drug response in multiple formats through either drug- or gene-specific pages. All data are freely available for download either through gene- or drug-specific pages, or as a whole through the download page.

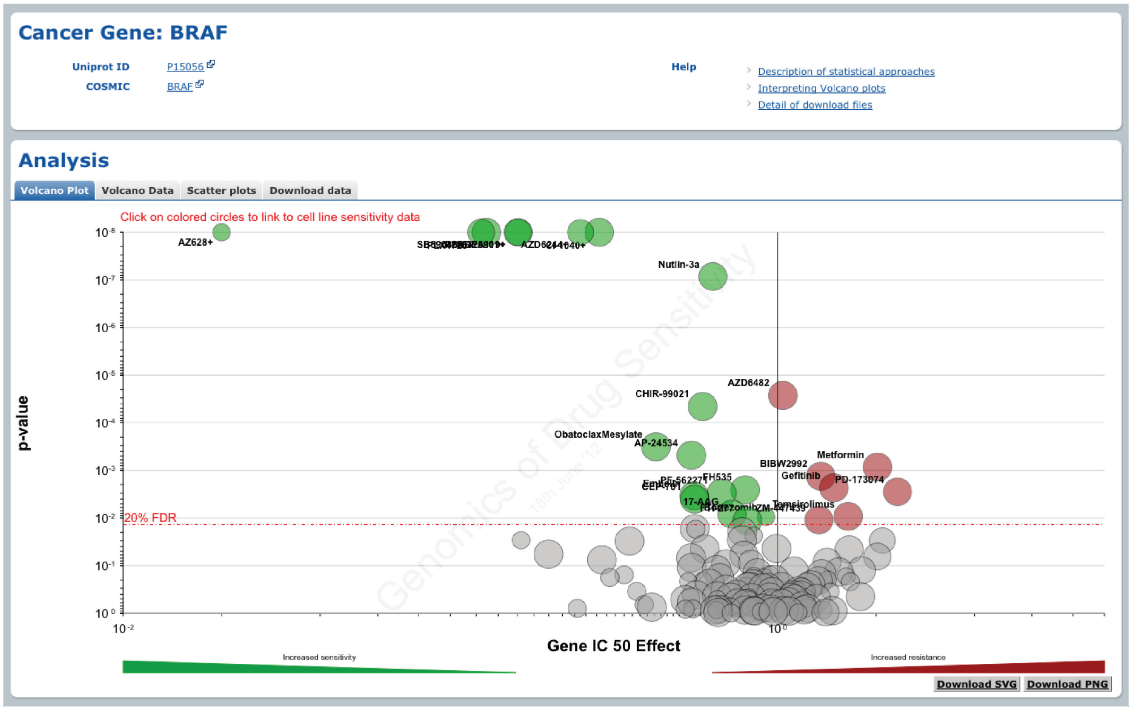

(a)
Drug : PLX4720
Help
Description of statistical approaches
Interpreting Volcano plots
Interpreting EN heatmaps
Detail of download files
Analysis of drug sensitivity data
Volcano Plot
Volcano Data
Elastic net
Scatter plots
Download data
p-value
10⁻⁸
10⁻⁷
10⁻⁶
10⁻⁵
10⁻⁴
10⁻³
10⁻²
10⁻¹
10⁰
BRAF+
RB1
BCR_ABL
APC
FLCN
20% FDR
Genomics of Drug Sensitivity
18th June '12
10⁻²
10⁰
10²
Drug IC 50 Effect
Increased sensitivity
Increased resistance
Download SVG
Download PNG
(b)
Analysis of drug sensitivity data
Volcano Plot
Volcano Data
Elastic net
Scatter plots
Download data
BRAF MUT
skin
BCL2A1
GYPC
DAAM2
IC 50
DU-4475
EoL-1-cell
C32
M14
A101D
G-361
HT-144
ACN
MEL-HO
SK-MEL-3
A375
MMAC-SF
BHT-101
K5
BV-173
RVH-421
HCC2218
WM-115
SK-MEL-28
COLO-679
NB10
NB69
SW1783
Becker
LS-513
MDA-MB-231
SK-LMS-1
TGW
SF539
TE-10
LNCaP-Clone-FGC
MFH-ino
ES8
T84
HCE-T
HCC38
LAN-6
LK-2
COLO-678
MG-63
-15
-10
-5
0
3
2
1
0
Expression Level
6
4
2
0
-2
IC 50 (natural log)

Targets BRAF
PubCHEM 24180719

Click on colored circles to link to cell line sensitivity data

Figure 2. Querying the GDSC database by compound name. Drug-specific pages demonstrate the effect of genomic features on cell line sensitivity to a particular drug. In this example, we show the effect of genomic features on sensitivity to the BRAF-inhibitor PLX4720. (**a**) A volcano plot representation of MANOVA results showing the magnitude (*x*-axis) and significance (*P*-value, log scale on inverted axis) for each cancer gene association. Each circle represents a single drug–gene interaction and the size is proportional to the number of mutant cell lines screened for each drug. For clarity, the *y*-axis is capped at $P = 1 \times 10^{-8}$ and a plus sign (+) next to a circle indicates that the *P*-value is smaller than this threshold. The dashed red line represents a Benjamini–Hochberg multiple testing correction for significance and only significant associations are coloured either green for drug sensitivity or red for resistance. (**b**) Elastic net analysis of genomic features associated with sensitivity to PLX4720. Features with negative effect size are associated with drug sensitivity and features with positive effect size are associated with drug resistance (all features are negative in this example). Mutation and tissue features are at the top of the heatmap to represent the presence (black) or absence (grey) of a mutation/tissue subtype. Below this are gene expression and copy number features with blue corresponding to lower expression or copy number, and red to indicate higher expression or copy number.

drug-specific volcano plot for the BRAF-inhibitor PLX4720 shows that mutations in the gene *BRAF* are significantly associated with sensitivity to this compound (Figure 2a). Conversely, the gene-specific volcano plot for *BRAF* shows that mutations in this gene are associated with sensitivity to multiple drugs including several different BRAF inhibitors (i.e. PLX4720, SB590885 and AZ628) (Figure 3). In both cases, the *x*-axis represents the magnitude of the effect of a gene–drug interaction on IC_{50} values across the cell lines screened and the *y*-axis is the significance of the interaction (*P*-value). By hovering over each circle, the following information is provided: genetic



Cancer Gene: BRAF

Uniprot ID P15056

COSMIC BRAF

Help
Description of statistical approaches
Interpreting Volcano plots
Detail of download files
Analysis
Volcano Plot
Volcano Data
Scatter plots
Download data
Click on colored circles to link to cell line sensitivity data
p-value
10^{-8}
10^{-7}
10^{-6}
10^{-5}
10^{-4}
10^{-3}
10^{-2}
10^{-1}
10^{0}
10^{-2}
10^{0}
AZ628+
Nutlin-3a
AZD6482
CHIR-99021
ObatoclaxMesylate
AP-24534
BIBW2992
Metformin
Gefitinib
PD-173074
FH535
CEP-701
17-AAG
Temsirolimus
20% FDR
Gene IC 50 Effect
Increased sensitivity
Increased resistance
Download SVG
Download PNG

Figure 3. Querying the GDSC database by cancer gene. Gene-specific pages show how a cancer gene mutation affects response to many drugs. A volcano plot representation shows results of the MANOVA analysis for drug sensitivity associated with *BRAF* mutations.

event sample size (i.e. the number of cell lines screened with a specific mutation), effect size and *P*-value. By clicking on an individual circle, it is possible to link to a scatter plot of cell line IC_{50} values for this association (see below). The volcano data tab represents the volcano plot data as a sortable table. Three buttons at the top of the table allow the download of the table in .csv, .tab or .xlsx file format.

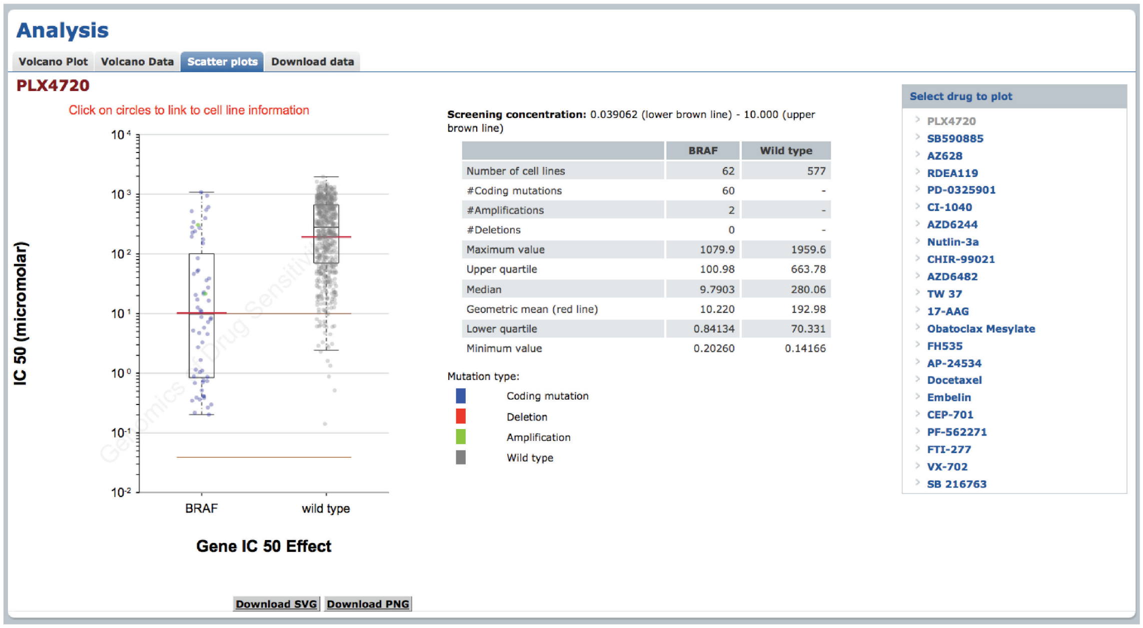
Similarly, the elastic net tab contains a graphical representation of results from the elastic net analysis of drug sensitivity (Figure 2b). For effective visualization, a maximum of 10 significant features associated with drug response are shown. These may include tissue type, mutations in cancer genes, expression levels and gene copy number. Each graphic contains three elements: a bar plot of effect size for significant features (right-hand side), a heatmap of genomic features (central panel) and a second heatmap of IC_{50} values for the 20 least and most sensitive cell lines (bottom). For example, the elastic net analysis for BRAF-inhibitor PLX4720 identified mutations in the *BRAF* gene, the tissue-type skin, as well as several transcriptional features (*BCL2A1*, *GYPC* and *DAAM2*) as associated with drug sensitivity (Figure 2b). Unlike the MANOVA analysis, gene-specific correlations for the elastic net analysis are not represented since the EN describes how multiple genes affect drug sensitivity together.

The 'Scatter plots' tab shows a plot of cell line IC_{50} values to a drug. IC_{50} values are split into two populations according to a cell lines mutational status for a given gene that is significantly associated with the drug response (Figure 4). In the example provided, cell lines with a *BRAF* mutation are on average more sensitive to PLX4720 compared with *BRAF* wild-type cell lines (Figure 4). The table in the middle shows the statistics for the plot including sample size, and the mean and median IC_{50} values for the two populations (mutated or wild type). Additional functionality includes the ability of users to select which drugs (or genes from the drug-specific pages) to plot. It is possible to link directly to relevant scatter plots by clicking on circles within the volcano plot pages. Furthermore, by clicking on circles within scatter plots, cell line IC_{50} values are directly linked to the COSMIC database facilitating integration of drug sensitivity data with detailed cell line information such a tissue type, tumour histology and a description of cell line origin and genotype.

Download data

As the emphasis of the website is on the graphical representation of data both the volcano and scatter plots are downloadable as either .png or .svg files. In addition, the raw data are available to download in either .csv or .xlsx format. As described below, it is possible to download the data for a specific drug or gene on their associated pages, or to download the data from all of our analyses in a series of large spreadsheets.

On the drug page for a specific compound, the available downloads include (i) sensitivity data for the drug (a table of cell line IC_{50} values); (ii) genomic alterations in cell lines; (iii) genomic correlations with MANOVA and (iv) elastic net analysis of drug sensitivity. On a gene page a single data download is available, containing the


Analysis
Volcano Plot
Volcano Data
Scatter plots
Download data
PLX4720
Click on circles to link to cell line information
IC 50 (micromolar)
10^4
10^3
10^2
10^1
10^0
10^{-1}
10^{-2}
BRAF
wild type
Gene IC 50 Effect
Genomics of Drug Sensitivity
Download SVG
Download PNG
Mutation type:
Coding mutation
Deletion
Amplification
Wild type
Select drug to plot
PLX4720
SB590885
AZ628
RDEA119
PD-0325901
CI-1040
AZD6244
Nutlin-3a
CHIR-99021
AZD6482
TW 37
17-AAG
Obatoclax Mesylate
FH535
AP-24534
Docetaxel
Embelin
CEP-701
PF-562271
FTI-277
VX-702
SB 216763

Screening concentration: 0.039062 (lower brown line) - 10.000 (upper brown line)

	BRAF	Wild type
Number of cell lines	62	577
#Coding mutations	60	-
#Amplifications	2	-
#Deletions	0	-
Maximum value	1079.9	1959.6
Upper quartile	100.98	663.78
Median	9.7903	280.06
Geometric mean (red line)	10.220	192.98
Lower quartile	0.84134	70.331
Minimum value	0.20260	0.14166

Figure 4. Scatter plot of cell line IC_{50} values and the effect of a cancer mutation. A scatter plot of cell line IC_{50} values for *BRAF*-mutated versus wild-type cell lines following drugging with PLX4720. Each circle represents the IC_{50} value for an individual cell line plotted on a logarithmic scale and the red line is the geometric mean of the population. Cell lines are colour coded to indicate whether the mutation is a coding mutation detected by sequencing, or an amplification or deletion detected by copy number analysis. The lower and upper brown lines indicate the minimum and maximum concentration in micro-molar of drug used for screening. Super-imposed on the scatter plot is a box-and-whisker plot showing the median, interquartile ranges and max and min for each plot. The central panel contains statistics for the plot and the right-hand table allows users to select which drug data to plot.

MANOVA correlation for how a gene correlates with drug response across the entire panel of compounds.

Rather than downloading drug- or gene-specific data, drug sensitivity and genomic datasets can also be directly downloaded as a whole through the 'Downloads' page. This can be directly accessed from the header on each page. Downloadable files include (i) cell line tissue type, drug sensitivity and genomic data used for the MANOVA; (ii) the MANOVA results for all compounds; (iii) A tissue-specific ANOVA to examine the effect of tissue type on drug response; (iv) the elastic net results for all compounds; (v) cell line genomic and transcriptional data used for elastic net analysis and (vi) a continuously updated list of cancer cell lines in our collection. Please note that some of these files contain a large number of columns and data will be lost if files are opened in Excel 2003 or earlier versions because the worksheet size is limited to 256 columns. The 'Downloads' page also provides access to archive files of previous data releases.

FUTURE WORK

The GDSC database will expand significantly in coming years as the size and complexity of datasets increase. The database currently contains data for ~75 000 experiments across 138 drugs and the amount of drug sensitivity data is expected to increase in size 2–3-fold within the next 2 years and even further in the future. This will include drug sensitivity data for many new cell lines to bring the total number to >1000 lines, and the inclusion of data for hundreds of newly screened anticancer drugs. Collectively, this will expand the number of different cancer subtypes and genotypes represented within the cell line collection, as well as the number of different drug targets interrogated by screening compounds.

Additional developments will see the further genomic characterization of the cell line collection to increase its utility as a resource. Notably, this will include whole-exome sequencing of all ~22 000 coding genes across the entire collection. Whole genome SNP6.0 copy number data currently include ~750 cell lines and this will be expanded to include the entire cell line collection. Similarly, basal transcriptional data are currently being updated to include the entire cell line collection using the latest Affymetrix human genome U219 mRNA expression array. These new genomic datasets, together with our expanding drug sensitivity datasets, will be incorporated into our analytical models to enhance our ability to identify therapeutic biomarkers predictive of drug response.

Large numbers of primary tumours across different cancer types are being extensively genomically characterized by systematic efforts such as the International Cancer Genomics Consortium. This will give us profound insights into the molecular taxonomy of cancer and, for the first time, enable us to directly assess the genomic similarity of our cell line models to primary tumours. Based on these comparisons, we will refine and expand the cell line collection to ensure that they are as representative as possible of primary tumours. Similarly, it is increasingly clear that patient responses to therapy can be highly variable even within pre-selected populations. We intend to systematically interrogate mechanisms of intrinsic resistance to clinical drugs by enriching the cell line collection with suitable *in vitro* models. An additional development will be an increased emphasis on drug sensitivity screening of low passage primary cultures derived from patient tumours. Collectively, these developments should further enhance our ability to model drug response and help address the fundamental question of intrinsic resistance to clinical cancer drugs.

The expansion of the database will likely require the development of new analytical tools to identify genomic biomarkers of drug response. This will involve the development of new statistical methods to interrogate the data and is likely to include the incorporation of pathway and tissue-based analyses of drug sensitivity. Similarly, the expansion will require new tools for mining and visualizing increasingly sophisticated and complex analyses.

DISCUSSION

Here, we have presented the GDSC database as a new resource for therapeutic biomarker discovery in cancer cells. Key features of the database include the largest resource of cell line anticancer drug sensitivity data publically available. In addition, the GDSC database integrates large genomic datasets with drug sensitivity information to identify putative therapeutic biomarkers for further preclinical validation. These data are presented using simple graphical representations and all data are freely available for download. The GDSC database will undergo significant expansion in coming years as drug sensitivity and genomic datasets increase in size and complexity.

The ultimate aim of the GDSC database is to facilitate the development of new cancer therapies through the preclinical identification of therapeutic biomarkers. The current approach to development of new cancer therapies is difficult, costly and time-consuming. For example, the time taken to develop a new drug is often >10 years and costs frequently exceed 1 billion US dollars. Moreover, despite this substantial investment, the level of attrition is very high with the majority of new drugs failing during clinical trials (estimated to be between an 80–95% failure rate) due to lack of efficacy or unacceptable toxicity (11). The preclinical identification of therapeutic biomarkers could significantly improve the design and ultimate success of clinical trials by permitting smaller, quicker and less costly trials in molecularly stratified patient populations most likely to benefit from treatment. By facilitating the preclinical identification of putative therapeutic biomarkers, the GDSC database is a valuable resource to enable the development of new rationally designed cancer therapeutic strategies incorporating molecular biomarkers.

ACKNOWLEGEMENTS

We would like to thank the members of the GDSC team for advice and discussion. We also thank Prasad Gunasekaran, Mingming Jia and Paul Bevan for their technical assistance.

FUNDING

The Wellcome Trust [086357 to M.R.S., P.A.F., D.A.H.]; Royal College of Physicians and Surgeons of Glasgow/Cancer Research UK Clinician Scientist Fellowship (to U.M.). Funding for open access charge: Wellcome Trust.

Conflict of interest statement. None declared.

REFERENCES

1. Druker,B.J., Guilhot,F., O'Brien,S.G., Gathmann,I., Kantarjian,H., Gattermann,N., Deininger,M.W., Silver,R.T., Goldman,J.M., Stone,R.M. *et al.* (2006) Five-year follow-up of patients receiving imatinib for chronic myeloid leukemia. *N. Engl. J. Med.*, **355**, 2408–2417.
2. Chapman,P.B., Hauschild,A., Robert,C., Haanen,J.B., Ascierto,P., Larkin,J., Dummer,R., Garbe,C., Testori,A., Maio,M. *et al.* (2011) Improved survival with vemurafenib in melanoma with BRAF V600E mutation. *N. Engl. J. Med.*, **364**, 2507–2516.
3. Kwak,E.L., Bang,Y.J., Camidge,D.R., Shaw,A.T., Solomon,B., Maki,R.G., Ou,S.H., Dezube,B.J., Janne,P.A., Costa,D.B. *et al.* (2010) Anaplastic lymphoma kinase inhibition in non-small-cell lung cancer. *N. Engl. J. Med.*, **363**, 1693–1703.
4. Stratton,M.R. (2011) Exploring the genomes of cancer cells: progress and promise. *Science*, **331**, 1553–1558.
5. Barretina,J., Caponigro,G., Stransky,N., Venkatesan,K., Margolin,A.A., Kim,S., Wilson,C.J., Lehar,J., Kryukov,G.V., Sonkin,D. *et al.* (2012) The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, **483**, 603–607.
6. McDermott,U., Sharma,S.V., Dowell,L., Greninger,P., Montagut,C., Lamb,J., Archibald,H., Raudales,R., Tam,A., Lee,D. *et al.* (2007) Identification of genotype-correlated sensitivity to selective kinase inhibitors by using high-throughput tumor cell line profiling. *Proc. Natl Acad. Sci. USA*, **104**, 19936–19941.
7. Garnett,M.J., Edelman,E.J., Heidorn,S.J., Greenman,C.D., Dastur,A., Lau,K.W., Greninger,P., Thompson,I.R., Luo,X., Soares,J. *et al.* (2012) Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, **483**, 570–575.
8. Forbes,S.A., Bindal,N., Bamford,S., Cole,C., Kok,C.Y., Beare,D., Jia,M., Shepherd,R., Leung,K., Menzies,A. *et al.* (2011) COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res.*, **39**, D945–D950.
9. Bolton,E.E., Wang,Y., Thiessen,P.A. and Bryant,S.H. (2008) PubChem: Integrated Platform of Small Molecules and Biological Activities. In: Ralph,A.W. and David,C.S. (eds), *Annual Reports in Computational Chemistry*, 4th edn. Elsevier, pp. 217–241.
10. UniProt Consortium. (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **40**, D71–D75.
11. Walker,I. and Newell,H. (2009) Do molecularly targeted agents in oncology have reduced attrition rates? *Nat. Rev. Drug Discov.*, **8**, 15–16.

View publication stats

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/23253235

Group A Human Rotavirus Genomics: Evidence that Gene Constellations Are Influenced by Viral Protein Interactions

Article *in* Journal of Virology · January 2008

DOI: 10.1128/JVI.01402-08 · Source: PubMed

CITATIONS
170

READS
227

6 authors, including:

Sarah M Mcdonald
Wake Forest University
145 PUBLICATIONS **4,314** CITATIONS
SEE PROFILE

Mario Barro
Sanofi Pasteur MSD
122 PUBLICATIONS **1,283** CITATIONS
SEE PROFILE

Tamara Bar-magen
University of Helsinki
21 PUBLICATIONS **839** CITATIONS
SEE PROFILE

John T Patton
Indiana University Bloomington
327 PUBLICATIONS **12,720** CITATIONS
SEE PROFILE

All content following this page was uploaded by John T Patton on 03 June 2014.

The user has requested enhancement of the downloaded file.

JOURNAL OF VIROLOGY, Nov. 2008, p. 11106–11116 Vol. 82, No. 22
0022-538X/08/$08.00+0 doi:10.1128/JVI.01402-08

Group A Human Rotavirus Genomics: Evidence that Gene Constellations Are Influenced by Viral Protein Interactions[▽][†]

Erica M. Heiman,[‡][§] Sarah M. McDonald,[‡] Mario Barro,[¶] Zenobia F. Taraporewala,
Tamara Bar-Magen,[‖] and John T. Patton[*]

*Laboratory of Infectious Diseases, National Institute of Allergy and Infectious Diseases, National Institutes of Health,
50 South Drive, MSC 8026, Room 6314, Bethesda, Maryland 20892-8026*

Received 5 July 2008/Accepted 3 September 2008

Group A human rotaviruses (HRVs) are the major cause of severe viral gastroenteritis in infants and young children. To gain insight into the level of genetic variation among HRVs, we determined the genome sequences for 10 strains belonging to different VP7 serotypes (G types). The HRVs chosen for this study, D, DS-1, P, ST3, IAL28, Se584, 69M, WI61, A64, and L26, were isolated from infected persons and adapted to cell culture to use as serotype references. Our sequencing results revealed that most of the individual proteins from each HRV belong to one of three genotypes (1, 2, or 3) based on their similarities to proteins of genogroup strains (Wa, DS-1, or AU-1, respectively). Strains D, P, ST3, IAL28, and WI61 encode genotype 1 (Wa-like) proteins, whereas strains DS-1 and 69M encode genotype 2 (DS-1-like) proteins. Of the 10 HRVs sequenced, 3 of them (Se584, A64, and L26) encode proteins belonging to more than one genotype, indicating that they are intergenogroup reassortants. We used amino acid sequence alignments to identify residues that distinguish proteins belonging to HRV genotype 1, 2, or 3. These genotype-specific changes cluster in definitive regions within each viral protein, many of which are sites of known protein-protein interactions. For the intermediate viral capsid protein (VP6), the changes map onto the atomic structure at the VP2-VP6, VP4-VP6, and VP7-VP6 interfaces. The results of this study provide evidence that group A HRV gene constellations exist and may be influenced by interactions among viral proteins during replication.

Group A human rotaviruses (HRVs) cause severe gastroenteritis in children and are responsible for up to 600,000 deaths each year worldwide (37). The rotavirus genome consists of 11 segments of double-stranded RNA contained within a triple-layered capsid (39). In general, each gene segment encodes a single polypeptide, allowing rotaviruses to express five or six nonstructural proteins (NSP1 to NSP5 and sometimes NSP6) and six capsid proteins (VP1 to VP4, VP6, and VP7) (12). The inner layer of the virion is composed of the core lattice protein (VP2) and encases the viral RNA-dependent RNA polymerase (VP1), the RNA capping enzyme (VP3), and genomic RNA (39). The intermediate layer is made entirely of VP6, which is the most recognized group- and subgroup (SG)-specific rotavirus antigen used for classification purposes (12, 23). Comprising the outermost layer of the rotavirus particle, the capsid proteins VP7 and VP4 are involved in virus entry into the cell and elicit neutralizing antibodies during infection of a host (12, 23). Serotypes defined by these outer capsid proteins are referred to as G types (VP7) and P types (VP4) and, along with G/P genotypes, are a predominant method of characterizing individual group A rotavirus isolates (23, 30). To date, a total of 14 G types (19 G genotypes) and 14 P types (27 P genotypes) have been identified in nature (23, 30).

Molecular epidemiology studies suggest that HRV G-type strains G1 to G4 are the most prevalent and are often combined with P-type specificities P[4], P[6], and P[8] (13, 23). Nonetheless, other serotypes, some of which bear resemblance to animal rotaviruses, are emerging as important agents of human disease in certain geographical regions (11, 13). For example, recent reports indicate an increase in incidences of HRV serotypes G5 and G8 in developing countries in South America and Africa (25). Unusual HRV serotypes, such as G9 and G12, are being reported at high levels in many countries, including those that are industrialized (13, 43). Of the 19 known G genotypes, 11 of them have been isolated from patients hospitalized for gastroenteritis; genotypes G7 and G13 to G19 have so far been isolated only from animals (25). These epidemiological observations suggest that the circulating rotavirus population is diverse and evolving. However, aside from the genes encoding outer capsid proteins (VP7 and VP4), little is known about the genetic makeup of emerging and endemic HRV strains because very few full-genome sequences have been determined. The limited insight into nucleotide similarities among group A HRVs that we do have is based on the results of qualitative RNA hybridization assays. This approach assigns a viral isolate into one of three distinct genogroups when seven of its gene segments anneal with those of a genogroup prototype strain: Wa (G1P[8]), DS-1 (G2P[4]), or AU-1

---

[*] Corresponding author. Mailing address: Laboratory of Infectious Diseases, National Institute of Allergy and Infectious Diseases, National Institutes of Health, 50 South Drive, MSC 8026, Room 6314, Bethesda, MD 20892-8026. Phone: (301) 594-1615. Fax: (301) 496-8312. E-mail: jpatton@niaid.nih.gov.

[§] Present address: UC Berkeley-UCSF Joint Medical Program, 570 University Hall, Berkeley, CA 94720.

[¶] Present address: Global Vaccines, Inc., P.O. Box 14827, Research Triangle Park, NC.

[‖] Present address: McGill AIDS Centre, Lady Davis Institute, Jewish General Hospital, 3755 Cote-Ste-Catherine Road, Montreal, QC H3T 1E2, Canada.

[‡] These authors contributed equally to the manuscript.

[†] Supplemental material for this article may be found at http://jvi.asm.org/.

[▽] Published ahead of print on 10 September 2008.

(G3P[8]) (35). Of the three HRV genogroups, Wa and DS-1 are overwhelmingly prominent, whereas the AU-1 genogroup is extremely rare (34). Intragenogroup resemblances are often reflected in the antigenic properties of their VP6 proteins (SG-I versus SG-II) and the electrophoretic migration pattern of viral genome segments (short or long electropherotype) but not necessarily in their serotype specificities (34). For example, G3 HRV isolates have been categorized into each of the three genogroups (Wa, DS-1, and AU-1), suggesting that they are serotypically similar but genetically separate (34). The results of these hybridization studies reveal that the diversity of HRVs in nature is far greater and more complex than that indicated by serotyping alone. Still, genogroup analysis via RNA hybridization does not identify specific amino acid changes that distinguish proteins of one genogroup from another. With the increasing ease of genetic sequencing, studies aimed at comparative genomic analyses of HRV genogroups are warranted.

The factors that influence the emergence of new group A rotavirus serotypes and genogroups within the human population are not completely understood. Reassortment events, in which two distinct rotavirus strains exchange gene segments, are predicted to play a major role in generating diversity (13, 23). In support of this idea, strains with different serotypes and genogroup specificities have been found circulating together during a single rotavirus season, and coinfections of individuals with more than one rotavirus strain (some of animal origin) are reported with frequencies of up to 23% (14). As a result, rotaviruses are hypothesized to exist as a heterogeneous pool of reassortants that is under continuous extrinsic and intrinsic selection pressures (14). Maternal antibodies, previous infection, and vaccination are just some extrinsic, host-derived pressures that could lead to the emergence of novel HRV G/P types. Yet, there is also evidence suggesting that intrinsic selection pressures, such as the capacity of divergent rotavirus proteins to functionally interact during replication, are important in determining the emergence of reassortant strains. For instance, although the genes encoding VP7 and VP4 can segregate independently, certain combinations of G/P types seem preferred among naturally circulating HRVs; G1 is almost always associated with P[8], whereas G2 is commonly seen with P[4] (9). The existence of genogroups is also consistent with the idea that biases toward certain gene constellations among HRVs occur. In fact, independent segregation of genes between different genogroup strains, aside from those encoding VP7 and VP4, is uncommonly reported (49). It is possible that viral protein-protein interactions drive the intrinsic selection of HRV reassortant strains that can replicate better inside an infected cell. In other words, a Wa-like protein might function more effectively with another Wa-like protein than it would with a DS-1-like protein. To test this possibility, we need a more complete understanding of the genetic composition of proteins from group A HRVs belonging to different serotypes and genogroups.

The results presented in this report are part of a collaborative effort by our laboratory and others that resulted in the establishment of a classification system for all 11 group A rotavirus genome segments (30). Based on phylogenetic analyses of the nucleotide sequences, distinct genotypes were determined for each individual rotavirus segment. The results not only confirmed the G/P types of HRVs by use of appropriate percent identity cutoff values but also increased our understanding of the variation seen in the nine other rotavirus genes. The majority of analyzed HRV gene segments were assigned into genotypes that correlate with the previously described genogroup strains Wa (genotype 1), DS-1 (genotype 2), and AU-1 (genotype 3) (30). In this report, we present the history and detailed sequence characteristics of G-type laboratory reference strains D, DS-1, P, ST3, IAL28, Se584, 69M, WI61, A64, and L26. Our results are consistent with the idea that viral protein-protein interactions influence the extent of group A HRV diversity and the maintenance of preferred gene constellations.

MATERIALS AND METHODS

Cells and rotaviruses. Monkey kidney (MA104) cells were propagated in 199 medium supplemented to contain 10% fetal bovine serum, 1% penicillin-streptomycin, and 0.1% amphotericin B. Human cell culture-adapted rotavirus strains D (G1P[8]), DS-1 (G2P[4]), P (G3P[8]), ST3 (G4P[6]), IAL28 (G5P[8]), Se584 (G6P[9]), 69M (G8P[10]), WI61 (G9P[8]), A64 (G10P[14]), and L26 (G12P[4]) were generously provided by Y. Hoshino (NIH) and were used to infect MA104 cells after pretreatment with 10 μg/ml trypsin, as described previously (16). Infected cells were incubated in serum-free 199 medium for 24 to 72 h until cytopathic effect was visible.

RNA extraction, RT-PCR, and sequencing. Total cytoplasmic RNA was extracted from approximately 2×10^6 rotavirus-infected cells by use of Trizol-LS as described in the manufacturer's protocol (Invitrogen). RNA was denatured by incubation for 10 min at 94°C in a final concentration of 50% dimethyl sulfoxide. Denatured RNA (from approximately 1×10^4 infected cells) was then used as a template for reverse transcription (RT)-PCR to amplify cDNAs corresponding to regions of rotavirus gene 1, 2, 3, or 6 (encoding VP1, VP2, VP3, or VP6, respectively). For genes 5, 6, 9, 10, and 11, RT-PCR was performed using a SuperScript one-step RT-PCR system (Invitrogen). For genes 1 and 2 and genes 3, 4, 7, and 8, respectively, SuperScript II RT and Platinum *Taq* DNA polymerase (Invitrogen) were used according to the manufacturer's protocols. The PCR products were purified using a QIAquick PCR purification kit (Qiagen). In cases when nonspecific PCR bands were present, samples were electrophoresed in a 1% agarose gel, and specific cDNA products were excised and purified using a QIAquick gel extraction kit (Qiagen). The purified PCR products were then sequenced with an ABI Prism BigDye v3.1 terminator cycle sequencing kit (Applied Biosystems Group). The dye terminator was removed using Performa DTR (Edge Biosystems), and sequences were obtained with a 3100 genetic analyzer for genes 5, 8, and 11 or a 3730 DNA analyzer (both Applied Biosystems) for genes 1 to 4, 6, 7, 9, and 10.

For sequencing the 3′ ends of 69M and L26 gene 3 (approximately nucleotides 2155 to 2507) and the 3′ end of ST3 gene 7, the PCR products were ligated into the pGEM-T-Easy vector (Promega) and three independent clones were sequenced. For sequencing the 5′ ends of gene 4 for all of the viruses, PCR products were ligated into the pCR-TOPO-2.1 vector (Invitrogen). Strain A64 genes 10 and 11 were ligated to the pGEM-T-Easy vector and then cut with restriction enzymes BsmBI and MfeI (New England Biolabs), respectively, so that the duplicated genes could be sequenced accurately. Sequence files were analyzed and assembled using Sequencher 4.5 (Gene Codes Corporation).

The nucleotide sequences of the entire open reading frames (ORFs), as well as additional flanking sequences, were determined. Terminal primers were used for RT-PCR amplification, and therefore the extreme 3′ and 5′ termini were not determined de novo.

Amino acid alignments. Amino acid alignments were constructed with MacVector 8.1.2. (Accelrys) using ClustalW, BLOSUM Series, with the defaults set (open gap penalty of 10.0, extended gap penalty of 0.05, and delay divergence of 40%). GenBank accession numbers for the viruses used in this study are presented in Table 1.

RESULTS AND DISCUSSION

The nucleotide sequences of the gene segments for HRV G-type reference strains D, DS-1, P, ST3, IAL28, Se584, 69M, WI61, A64, and L26 were determined by automated sequencing. Table 2 summarizes pertinent features of these strains, as

TABLE 1. GenBank accession numbers

HRV strain	GenBank accession no.										
	VP1	VP2	VP3	VP4	VP6	VP7	NSP1	NSP2	NSP3	NSP4	NSP5/6
Wa	DQ490539	X14942	AY267335	L34161	K02086	M21843	L18943	L04534	X81434	AF093199	AF306494
KU	AB022765	AB022766	AB022767	AB222784	AB022768	D16343	AB022769	AB022770	AB022771	AB022772	AB022773
Dhaka 16-03	DQ492669	DQ492670	DQ492671	DQ492672	DQ492673	DQ492674	DQ492675	DQ492676	DQ492677	DQ492678	DQ492679
D	EF583021	EF583022	EF583023	EF672570	EF583024	EF672574	EF672571	EF672573	EF672572	EF672575	EF672576
DS-1	EF990691	EF990692	EF990693	EF672577	EF990694	EF672581	EF672578	EF672580	EF672579	EF672582	EF672583
TB-Chen	AY787653	AY787652	AY787654	AY787644	AY787645	AY787646	AY787647	AY787648	AY787649	AY787650	AY787651
P	EF583037	EF583038	EF583039	EF672598	EF583040	EF672602	EF672599	EF672601	EF672600	EF672603	EF672604
AU-1	DQ490533	DQ490536	DQ490537	D10970	DQ490538	D86271	D45244	DQ490534	DQ490535	D89873	AB008656
ST3	EF583045	EF583046	EF583047	EF672612	EF583048	EF672616	EF672613	EF672615	EF672614	EF672617	EF672618
IAL28	EF583029	EF583030	EF583031	EF672584	EF583032	EF672588	EF672585	EF672587	EF672586	EF672589	EF672590
Se584	EF583041	EF503842	EF583043	EF672605	EF583044	EF672609	EF672606	EF672608	EF672607	EF672610	EF672611
DRC86	DQ005125	DQ005124	DQ005123	DQ005122	DQ005121	DQ005120	DQ005119	DQ005118	DQ005117	DQ005116	DQ005115
DRC88	DQ005114	DQ005113	DQ005112	DQ005111	DQ005110	DQ005109	DQ005108	DQ005107	DQ005106	DQ005105	DQ005104
69M	EF576937	EF576914	EF576915	EF672556	EF576916	EF672560	EF672557	EF672559	EF672558	EF672561	EF672562
WI61	EF583049	EF583050	EF583051	EF672619	EF583052	EF672623	EF672620	EF672622	EF672621	EF672624	EF672625
A64	EF583017	EF583018	EF503819	EF672563	EF503820	EF672567	EF672564	EF672566	EF672565	EF672568	EF672569
L26	EF583033	EF583034	EF583035	EF672591	EF583036	EF672595	EF672592	EF672594	EF672593	EF672596	EF672597
Dhaka 12-03	DQ146660	DQ146661	DQ146662	DQ146663	DQ146664	DQ146665	DQ146666	DQ146667	DQ146668	DQ146669	DQ146670
Matlab13-03	DQ146671	DQ146672	DQ146673	DQ146674	DQ146675	DQ146676	DQ146677	DQ146678	DQ146679	DQ146680	DQ146681
N26-02	DQ146682	DQ146683	DQ146684	DQ146685	DQ146686	DQ146687	DQ146688	DQ146689	DQ146690	DQ146691	DQ146692
RV176-00	DQ490551	DQ490552	DQ490553	DQ490554	DQ490555	DQ490556	DQ490557	DQ490558	DQ490559	DQ490560	DQ490561
RV161-00	DQ490545	DQ490546	DQ490547	DQ490548	DQ490549	DQ490550	DQ490540	DQ490541	DQ490542	DQ490543	DQ490544
B4633-03	DQ146638	DQ146639	DQ146640	DQ146641	DQ146642	DQ146643	DQ146644	DQ146645	DQ146646	DQ146647	DQ146648
Dhaka 25-02	DQ146649	DQ146650	DQ146651	DQ146652	DQ146653	DQ146654	DQ146655	DQ146656	DQ146657	DQ146658	DQ146659
T152	DQ146699	DQ146700	DQ146701	AB077766	DQ146702	AB071404	AB097459	DQ146703	DQ146704	DQ146705	DQ146706

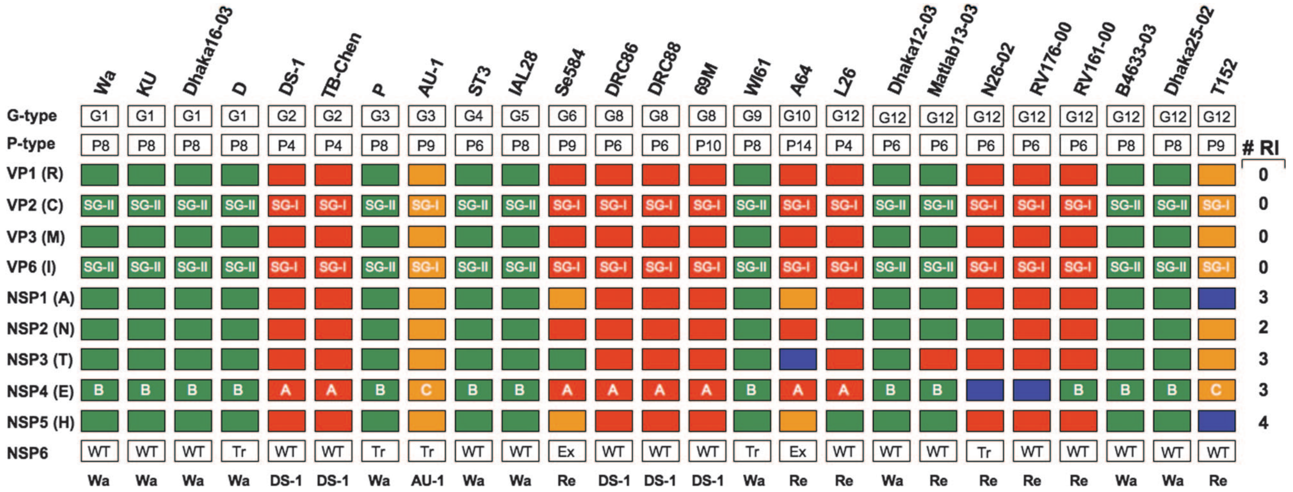
well as those of other group A HRVs whose genome sequences have been reported to date. With a few exceptions, the individual genes/proteins for the 10 reference strains were assigned to genotypes that correlate with previously described genogroup strains Wa (genotype 1), DS-1 (genotype 2), and AU-1 (genotype 3) (30) (Fig. 1). This classification system afforded the opportunity to identify HRVs that are either (i) pure genogroup strains, with all of their proteins classified into a single genotype, or (ii) intergenogroup reassortants, with proteins of more than one genotype.

TABLE 2. Group A human rotaviruses analyzed

HRV strain	G type	P type	Yr of isolation	Country of isolation
Wa	G1	P8	1974	United States
KU	G1	P8	1974	United States
Dhaka 16-03	G1	P8	2003	Bangladesh
D	G1	P8	1974	United States
DS-1	G2	P4	1976	United States
TB-Chen	G2	P4	1991	China
P	G3	P8	1974	United States
AU-1	G3	P9	1981–1983	Japan
ST3	G4	P6	1975	United Kingdom
IAL28	G5	P8	1992	Brazil
Se584	G6	P9	1998	United States
DRC86	G8	P6	2003	Democratic Republic of the Congo
DRC88	G8	P8	2003	Democratic Republic of the Congo
69M	G8	P10	1980	Indonesia
WI61	G9	P8	1983	United States
A64	G10	P14	1987	United Kingdom
L26	G12	P4	1987	Philippines
Dhaka 12-03	G12	P6	2003	Bangladesh
Matlab13-03	G12	P6	2003	Bangladesh
N26-02	G12	P6	2002	Bangladesh
RV176-00	G12	P6	2000	Bangladesh
RV161-00	G12	P6	2000	Bangladesh
B4633-03	G12	P8	2003	Belgium
Dhaka 25-02	G12	P8	2003	Bangladesh
T152	G12	P9	1998–1999	Thailand

Wa genogroup (genotype 1) strains D, P, ST3, IAL28, and WI61. The Wa strain is one of the best-studied HRVs, in large part because it was the first to be grown successfully in culture and because it causes acute diarrhea in gnotobiotic piglets (51). The virus was isolated in the United States in 1974, and as the prototype for the Wa genogroup, the virus exhibits a long electropherotype and VP6 SG-II and VP2 SG-II specificities (50) (Fig. 1 and Table 2). Using the aforementioned classification system, we found that 5 of the 10 viruses (D, P, ST3, IAL28, and WI61) are genetically more similar to Wa than to either the DS-1 or the AU-1 strain (30) (Fig. 1). With the exception of VP4 and VP7, each of the proteins encoded by these five viruses can be classified as genotype 1. The observation that these HRVs do not encode proteins similar to the DS-1 and AU-1 genogroups (genotypes 2 and 3, respectively) suggests that they are not intergenogroup reassortants but instead can be viewed as pure Wa genogroup strains. The Wa-like strains D, P, and ST3 have been characterized over the years because they represent three of the four most common HRV G types (G1, G3, and G4, respectively). The VP7-encoding genes from these strains (in addition to that of the G2 strain DS-1) have been incorporated into the live-attenuated, multivalent rotavirus vaccines RotaTeq and RotaShield (37). The D and P strains were originally isolated in 1974 in the United States (Children's Hospital, Washington, DC) from infants suffering from gastroenteritis (24, 50). The ST3 strain was isolated in 1975 from an asymptomatic newborn in England (St. Thomas Hospital, London) (6, 50). Despite the importance of the D, P, and ST3 strains, their genome sequences have not been determined previously.

While the D, P, and ST3 strains represent more commonly circulating HRV serotypes, IAL28 and WI61 reflect globally emerging serotypes G5 and G9, respectively (Table 2). Strain IAL28 was isolated in 1992 from a child with diarrhea in Brazil



	Wa	KU	Dhaka16-03	D	DS-1	TB-Chen	P	AU-1	ST3	IAL28	Se584	DRC86	DRC88	69M	WI61	A64	L26	Dhaka12-03	Matlab13-03	N26-02	RV176-00	RV161-00	B4633-03	Dhaka25-02	T152	# RI
G-type	G1	G1	G1	G1	G2	G2	G3	G3	G4	G5	G6	G8	G8	G8	G9	G10	G12	G12	G12	G12	G12	G12	G12	G12	G12	
P-type	P8	P8	P8	P8	P4	P4	P8	P9	P6	P8	P9	P6	P6	P10	P8	P14	P4	P6	P6	P6	P6	P6	P8	P8	P9	
VP1 (R)	🟩	🟩	🟩	🟩	🟥	🟥	🟩	🟧	🟩	🟩	🟥	🟥	🟥	🟥	🟩	🟥	🟥	🟩	🟩	🟥	🟥	🟥	🟩	🟩	🟧	0
VP2 (C)	🟩 SG-II	🟩 SG-II	🟩 SG-II	🟩 SG-II	🟥 SG-I	🟥 SG-I	🟩 SG-II	🟧 SG-I	🟩 SG-II	🟩 SG-II	🟥 SG-I	🟥 SG-I	🟥 SG-I	🟥 SG-I	🟩 SG-II	🟥 SG-I	🟥 SG-I	🟩 SG-II	🟩 SG-II	🟥 SG-I	🟥 SG-I	🟥 SG-I	🟩 SG-II	🟩 SG-II	🟧 SG-I	0
VP3 (M)	🟩	🟩	🟩	🟩	🟥	🟥	🟩	🟧	🟩	🟩	🟥	🟥	🟥	🟥	🟩	🟥	🟥	🟩	🟩	🟥	🟥	🟥	🟩	🟩	🟧	0
VP6 (I)	🟩 SG-II	🟩 SG-II	🟩 SG-II	🟩 SG-II	🟥 SG-I	🟥 SG-I	🟩 SG-II	🟧 SG-I	🟩 SG-II	🟩 SG-II	🟥 SG-I	🟥 SG-I	🟥 SG-I	🟥 SG-I	🟩 SG-II	🟥 SG-I	🟥 SG-I	🟩 SG-II	🟩 SG-II	🟥 SG-I	🟥 SG-I	🟥 SG-I	🟩 SG-II	🟩 SG-II	🟧 SG-I	0
NSP1 (A)	🟩	🟩	🟩	🟩	🟥	🟥	🟩	🟧	🟩	🟩	🟧	🟥	🟥	🟥	🟩	🟧	🟥	🟩	🟩	🟥	🟥	🟥	🟩	🟩	🟦	3
NSP2 (N)	🟩	🟩	🟩	🟩	🟥	🟥	🟩	🟧	🟩	🟩	🟥	🟥	🟥	🟥	🟩	🟥	🟩	🟩	🟩	🟩	🟥	🟥	🟩	🟩	🟧	2
NSP3 (T)	🟩	🟩	🟩	🟩	🟥	🟥	🟩	🟧	🟩	🟩	🟩	🟥	🟥	🟥	🟩	🟦	🟥	🟩	🟥	🟥	🟥	🟥	🟩	🟩	🟧	3
NSP4 (E)	🟩 B	🟩 B	🟩 B	🟩 B	🟥 A	🟥 A	🟩 B	🟧 C	🟩 B	🟩 B	🟥 A	🟥 A	🟥 A	🟥 A	🟩 B	🟥 A	🟥 A	🟩 B	🟩 B	🟦	🟦	🟩 B	🟩 B	🟩 B	🟧 C	3
NSP5 (H)	🟩	🟩	🟩	🟩	🟥	🟥	🟩	🟧	🟩	🟩	🟧	🟥	🟥	🟥	🟩	🟧	🟩	🟩	🟩	🟥	🟥	🟥	🟩	🟩	🟦	4
NSP6	WT	WT	WT	Tr	WT	WT	Tr	Tr	WT	WT	Ex	WT	WT	WT	Tr	Ex	WT	WT	WT	Tr	WT	WT	WT	WT	WT	
	Wa	Wa	Wa	Wa	DS-1	DS-1	Wa	AU-1	Wa	Wa	Re	DS-1	DS-1	DS-1	Wa	Re	Re	Wa	Re	Re	Re	Re	Wa	Wa	Re	

FIG. 1. Summary of HRV protein genotypes, genogroups, and reassortment incidences. Individual HRV strains analyzed in this study are listed at the top, with their corresponding genes/proteins aligned underneath. The VP7 and VP4 genotypes (G/P types) of the strains are listed. Colors indicate the genotype of each gene/protein, as described previously (30): green for Wa-like genotype 1, red for DS-1-like genotype 2, orange for AU-1-like genotype 3, and blue for animal RV-like genotypes. The single-letter abbreviation for each protein is indicated in parentheses (30). VP6 and VP2 SG specificities and NSP4 genotypes (A, B, and C) are classified according to previous studies (17, 26, 32). The NSP6 ORFs were found to be identical to that of the Wa strain (WT), truncated (Tr), or extended (Ex) (also see Fig. S3 in the supplemental material). Below each genome, the HRVs are classified as pure Wa genogroup strains (Wa), pure DS-1 genogroup strains (DS-1), a pure AU-1 strain (AU-1), or intergenogroup/interspecies reassortants (Re). The numbers of reassortment incidences (# RI) detected in this study for genes encoding VP1 to -3, VP6, and NSP1 to -5 are listed on the right.

and was initially shown to have dual G5-G11 antigenic determinants; this virus was later classified as a G5 strain based on sequence analysis (30, 48). Because G5 and G11 serotypes are commonly seen in pigs, IAL28 might represent an interspecies reassortant, with its VP7 gene coming from a porcine virus and its 10 other genes coming from a Wa genogroup HRV (48). Strain WI61 was recovered in 1983 from a child with gastroenteritis in the United States (Children's Hospital, Philadelphia, PA) (7). At the time of its isolation, only G1 to G4 serotypes had been identified in humans, and therefore WI61 represented the fifth G type to be discovered. With the recognition that the incidence of G9 HRV infections is steadily increasing, WI61 becomes a particularly important reference strain (44). In addition to those of D, P, ST3, IAL28, and WI61, the full-genome sequences of other Wa genogroup HRVs have been determined, including two G1P[8] strains (KU and Dhaka 16-03) and several G12 isolates: Dhaka 12-03 (G12P[6]), B4633-03 (G12P[8]), and Dhaka 25-02 (G12P[8]) (Table 1) (43, 50). Our analyses suggest that despite the differences in serotype specificities (i.e., G1, G3, G4, G5, G9, and G12), these HRVs are genetically similar viruses (Fig. 1). Thus, pure Wa genogroup viruses can bear a large number of VP7 serotypes.

DS-1 genogroup (genotype 2) strains DS-1 and 69M. Strain DS-1 is the well-known prototype for its genogroup and was isolated in 1976 in the United States (Children's Hospital, Washington, DC) from an infant with diarrhea (45, 50) (Table 2). Like many other strains belonging to the DS-1 genogroup, the virus has VP6 SG-I and VP2 SG-I specificities and a short electropherotype, the latter due to an extended gene 11 3′ untranslated region (UTR) (28). In addition to Wa-like strains D, P, and ST3, the VP7 gene of DS-1 represents one of the most predominant HRV serotypes (G2) and is included as a component of the tetravalent rotavirus vaccines. This is the first report of the full-genome sequence for this important vaccine-related strain.

Of the HRVs sequenced in this study, 69M is the only strain besides DS-1 that is classified as a pure member of the DS-1 genogroup (Fig. 1). 69M was isolated in Indonesia in 1980 from a child with diarrhea and is unusual in that, instead of the short electropherotype of most DS-1-like strains, 69M has a super-short electropherotype (19, 28, 29). The super-short migration pattern of the 69M genome is due to its gene 11 3′ UTR being twice the length of the already extended DS-1 gene 11 3′ UTR (28). Since 69M is a pure member of the DS-1 genogroup, all of the individual proteins encoded by 69M, with the exception of VP4 and VP7, can be classified as genotype 2. Other DS-1-like HRVs of which genome sequences are known include the Chinese isolate TB-Chen (G2P[4]) and the DS-1-like isolates DRC86 (G8P[6]) and DRC87 (G8P[8]), which are from the Democratic Republic of the Congo (5, 31). These results indicate that DS-1 genogroup HRVs can display VP7 proteins from diverse G types (e.g., G2 and G8), although seemingly not as diverse as the G types that are associated with Wa genogroup strains.

Intergenogroup reassortant strains Se584, A64, and L26. We found that three of the reference strains we sequenced, namely, Se584, A64, and L26, show proteins originating from more than one genotype. These three viruses encode mostly genotype 2 (DS-1-like) proteins but have a few proteins that can be classified as either genotype 1 or genotype 3 (Wa-like or AU-1-like, respectively). Strain AU-1 was isolated in Japan during the years 1981 to 1983, and based on the lack of RNA hybridization with the Wa or DS-1 probe, AU-1 became the prototype strain for a third HRV genogroup (34, 36). However, this genogroup is very rare, and with the exception of the prototype strain, no pure AU-1 HRVs have been isolated to date. The only other AU-1 genogroup strain recovered from a human is strain T152 (G12P[9]), which was discovered in Thailand and encodes only seven genotype 3 proteins (Fig. 1) (30, 42, 43). Interestingly, T152 expresses two proteins (NSP1 and NSP5) more closely related to animal rotaviruses, suggesting that this AU-1-like virus may represent an interspecies reassortant (30, 43). For HRVs of which full-genome sequences are known, we find that genotype 3 proteins tend to appear in reassortant strains with predominantly genotype 2 (DS-1-like) backgrounds (Fig. 1).

Of the 10 HRV reference strains, only strains Se584 and A64 encode proteins that can be classified as genotype 3 (30). Strain Se584 was isolated in 1998 from a 7-year-old child as part of a rotavirus surveillance project in the United States (18). The occurrence of this G6 serotype strain, especially in an older child, suggests that Se584 is an unusual isolate. Previous RNA hybridization studies of this virus showed that the AU-1 genogroup probes annealed to three Se584 gene segments and that no gene segments were recognized by the Wa or DS-1 probe (18). Our sequencing results contradict these RNA hybridization data; we found that Se584 encodes six genotype 2 proteins (VP1, VP2, VP3, VP6, NSP2, and NSP4), one genotype 1 protein (NSP3), and two genotype 3 proteins (NSP1 and NSP5) (Fig. 1). The reason for the lack of detectable hybridization of Se584 gene segments with the Wa or DS-1 probe in the previous study is unknown.

Our sequencing results show that strain A64 is similar to Se584 and encodes six genotype 2 proteins (VP1, VP2, VP3, VP6, NSP2, and NSP4) and two genotype 3 proteins (NSP1 and NSP5) (Fig. 1). However, while Se584 possesses a genotype 1 NSP3, the A64 NSP3 protein is not phylogenetically related to any HRV but instead seems related to an animal rotavirus (i.e., the bovine-like genotype 6) (30) (Fig. 1). Strain A64 is an unusual HRV in that it was isolated from an immunocompromised child (20) and therefore grew under conditions with atypical selective pressures. In addition to being a reassortant, this virus contains rearrangements in genes 7, 10, and 11 that stem from head-to-tail sequence duplications (2, 20). As shown in Fig. 2, all three sequence duplications initiate downstream of the ORF, thereby leaving genes 7, 10, and 11 capable of encoding full-length proteins (NSP3, NSP4, and NSP5/6, respectively). The sequence duplication in gene 11 introduces a second ORF for NSP6 in the viral RNA; yet, given its downstream position, it is not clear whether this ORF is capable of being translated (Fig. 2). Our description of the duplication of the A64 gene 10 is in agreement with that of a previous report (20). For A64 genes 7 and 11, this is the first description of the sequence duplications.

Strain L26 also represents an intergenogroup reassortant, in this case formed by genetic exchange of genes from viruses belonging to the Wa and DS-1 genogroups. The L26 strain is identified as a member of the DS-1 genogroup, since seven of its genes encode genotype 2 proteins, with two encoding genotype 1 proteins (NSP2 and NSP5) (Fig. 1). Originally isolated from a sick infant in the Philippines in 1987, L26 is a representative of the globally emerging G12 serotype (46). This strain has a long electropherotype but shows VP2 and VP6 SG-I specificities, an unusual combination for a DS-1 genogroup strain.

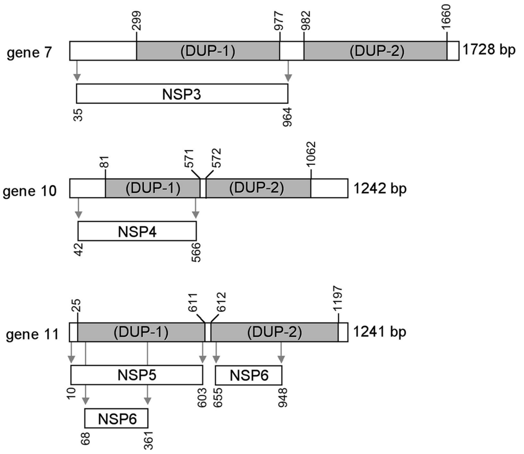


FIG. 2. Schematic of intragenic sequence rearrangements for A64 genes 7, 10, and 11. Each aberrant RNA segment contains a single head-to-tail sequence duplication (DUP-1 or DUP-2). The duplication initiates downstream of the ORF for each gene, thereby allowing A64 to express full-length NSP3, NSP4, and NSP5/6 proteins, respectively. The gene 11 rearrangement introduces a second NSP6 ORF, one that is positioned downstream of the NSP5 ORF.

Other HRVs that are known intergenogroup reassortants include several G12P[6] strains: Matlab13-03, N26-02, RV176-00, and RV161-00 (43). Matlab13-03 is predicted to belong to the Wa genogroup but encodes a genotype 2 NSP3 (Fig. 1). Strains N26-02, RV176-00, and RV161-00 can be described as DS-1-like, because the majority of their proteins are related to this genogroup (Fig. 1). N26-02 contains a genotype 1 NSP2 and an NSP4 from an animal rotavirus (genotype 6) (30) (Fig. 1). RV176-00 and RV161-00 encode NSP4 proteins that are genotype 1 and animal-like (genotype 6), respectively (30) (Fig. 1).

The fact that the HRV reassortants described above are viable demonstrates that the proteins encoded by the reassorted genes functionally replace the proteins encoded by the parental genes. Nonetheless, it is interesting that certain combinations of genes seem to exist, even among the reassortant strains. For example, the genes encoding VP1, VP2, VP3, and VP6 for each individual strain invariably belong to the same genotype. Additionally, we found that most of the reassortment events were associated with serotypes that are rarely seen in humans or that otherwise represented atypical infections (e.g., immunocompromised child or gastroenteritis in a 7-year-old) and that the predominant HRV serotypes (G1 to G4) showed no evidence of genetic exchange. Thus, these observations raise several important questions regarding possible intrinsic selection pressures influencing rotavirus diversity. Do rotaviruses have preferred gene constellations? Can proteins of different genotypes functionally substitute for each other in all cases? What are the precise amino acid changes that determine whether a protein belongs to genotype 1, 2, or 3? Do these changes cluster at sites of viral protein-protein interactions?

Correlating genotype-specific changes with protein interaction sites. To characterize the level of variation among HRV proteins belonging to genotype 1, 2, or 3, we first calculated the average intergenotype and intragenotype percent identities. In

TABLE 3. Intragenotype and intergenotype percent identities

Protein	% Identity (avg ± SD)[a] for genotypes:					
	1 vs 1	2 vs 2	3 vs 3	1 vs 2	1 vs 3	2 vs 3
VP1	98.2 ± 1.1	97.9 ± 0.9	98.3*	89.9 ± 0.4	91.3 ± 0.3	94.8 ± 0.3
VP2	97.1 ± 1.0	97.7 ± 1.8	98.8*	89.1 ± 0.8	91.5 ± 0.6	93.8 ± 0.4
VP3	96.6 ± 1.3	94.4 ± 3.6	93.2*	81.3 ± 0.6	83.4 ± 0.6	83.7 ± 0.4
VP6	98.7 ± 0.9	98.7 ± 0.8	97.5*	92.1 ± 0.4	91.8 ± 0.3	94.9 ± 1.0
NSP1	88.1 ± 5.8	96.6 ± 2.9	89.2 ± 5.1	68.9 ± 0.8	57.6 ± 1.2	58.5 ± 0.7
NSP2	95.9 ± 2.1	95.6 ± 1.7	95.0*	89.8 ± 0.8	89.3 ± 0.8	89.0 ± 0.8
NSP3	95.4 ± 2.7	97.5 ± 1.7	95.5*	82.0 ± 0.9	83.1 ± 0.8	88.8 ± 0.5
NSP4	95.9 ± 1.8	93.7 ± 1.9	94.3*	84.1 ± 1.2	84.1 ± 0.9	84.0 ± 1.2
NSP5	96.2 ± 2.3	96.1 ± 3.1	95.8 ± 1.1	83.0 ± 1.4	90.0 ± 0.9	86.4 ± 1.7

[a] *, single value; no standard deviation given.

this analysis, we utilized amino acid sequence information for the 25 HRVs for which full-genome sequences are available. We note that, due to the lack of availability of genotype 3 sequences, we are more confident regarding changes differentiating genotype 1 and 2 proteins. As shown in Table 3, proteins belonging to the same genotype (excluding NSP1) have a high degree of sequence identity (93.2 to 98.8%). In contrast, the sequence identity of proteins belonging to different genotypes is at least 5% lower. Thus, on average, proteins of different genotypes will differ by 5 or more amino acids per 100 from proteins of the same genotype, suggesting a greater evolutionary distance between genotypes. While the amino acid identities for NSP1 proteins of the same genotype ranged from 88.1 to 96.6%, the identities for those of different genotypes were substantially lower (57.6 to 68.9%) (Table 3). These data indicate that the evolutionary distance between NSP1 proteins of different genotypes vastly exceeds that of the other viral proteins analyzed (see below). To identify the precise amino acid residues that distinguish proteins from genotypes 1, 2, and 3, amino acid alignments were constructed (see Fig. S1 and Table S1 in the supplemental material). Residues in the alignments that differed depending on the genotype of the protein (vis-à-vis genotype-specific differences) were identified and mapped to defined regions or known structural or functional domains (Fig. 3). We found that for some of the proteins, genotype-specific changes often clustered in sites of known viral protein-protein interactions.

Interactions among VP1, VP2, VP3, and VP6. Proteins constituting the inner layers of the rotavirus virion (VP1, VP2, VP3, and VP6) segregate with a single genotype for each of the HRVs, suggesting that they are genetically linked (Fig. 1). Because these proteins are known to interact with each other, we hypothesize that the linkage is a reflection of amino acid changes driving more favorable intragenotype interactions. The core lattice protein (VP2) engages both the viral RNA-dependent RNA polymerase (VP1) and the RNA capping enzyme (VP3) via its internally located N-terminal domain (52). The interaction between the VP2 N terminus and VP1/VP3 is thought to be critical during rotavirus RNA synthesis (viral genome replication and transcription) (38). Additionally, the outer surface of the VP2 shell interacts with VP6 during the early stages of virion morphogenesis (12). The VP2-VP6 interaction is important for the formation of double-layered particles (DLPs) and also influences the transcriptional activity of the particle (12). The outer surface of the VP2 shell domain, which is accessible to interactions with VP6, exhibits 27 amino acid differences between genotypes 1 and 2 (32) (Fig. 3). However, we found that most of the genotype-specific differences in VP2 map to the N terminus of the protein and include dramatic sequence variations and insertions/deletions, which are predicted to influence VP1/VP3 binding (32) (see Fig. S2 in the supplemental material). For the viral polymerase VP1, we identified 66 amino acids that differ between genotypes 1 and 2 (Fig. 3). Although the changes are distributed throughout the enzyme, the bulk of them are surface exposed and concentrate at the N- and C-terminal domains and the finger subdomain; in contrast, the predicted catalytic domain of the VP1 is quite conserved (26a) (Fig. 3). To our surprise, VP3, the RNA methyltransferase/guanylyltransferase, shows more genotype-specific amino acid differences than any other inner capsid protein. We found 71 changes between genotypes 1 and 2 and 22 changes among genotypes 1, 2, and 3, many of which clustered in the central region and C terminus of the protein (Fig. 3). Based on what is seen for VP1, we predict that the currently undefined VP3 catalytic sites might be located in the more conserved regions of the protein. The regions of VP1 and VP3 that mediate interactions with each other and with the VP2 N terminus are unknown. It is possible that the identified genotype-specific changes could be sites of VP1/VP2/VP3 interactions.

Interactions among VP2, VP6, VP7, and VP4. Our sequence alignments indicate that VP6 contains relatively few genotype-specific amino acid differences (see Fig. S1 in the supplemental material). Those that are present tend to cluster toward the termini of the protein, with many in or close to known VP2 interactive domains (4) (Fig. 3). Mapping these differences onto the atomic structure of VP6 indicates that the majority of them (18 out of 23 amino acids) are surface exposed and that many are situated in domain B at the base of the VP6 trimer and near the VP2 contact site (Fig. 4). The placement of these residues raises the possibility that genotype-specific changes in VP6 could influence (i) the assembly of the VP6 protein layer onto VP2, (ii) the stability of the DLP, or (iii) the transcriptional activity of VP1 (4). VP6 also engages the outer capsid proteins VP7 and VP4 during formation of infectious triple-layered particles (12). The sites on VP6 where VP7 and VP4 bind have been described previously and can be visualized on the atomic structure of the trimer (27) (Fig. 4). VP6 contains several genotype-specific changes near the VP7-binding site. Interestingly, several of these residues comprise the VP6 SG-

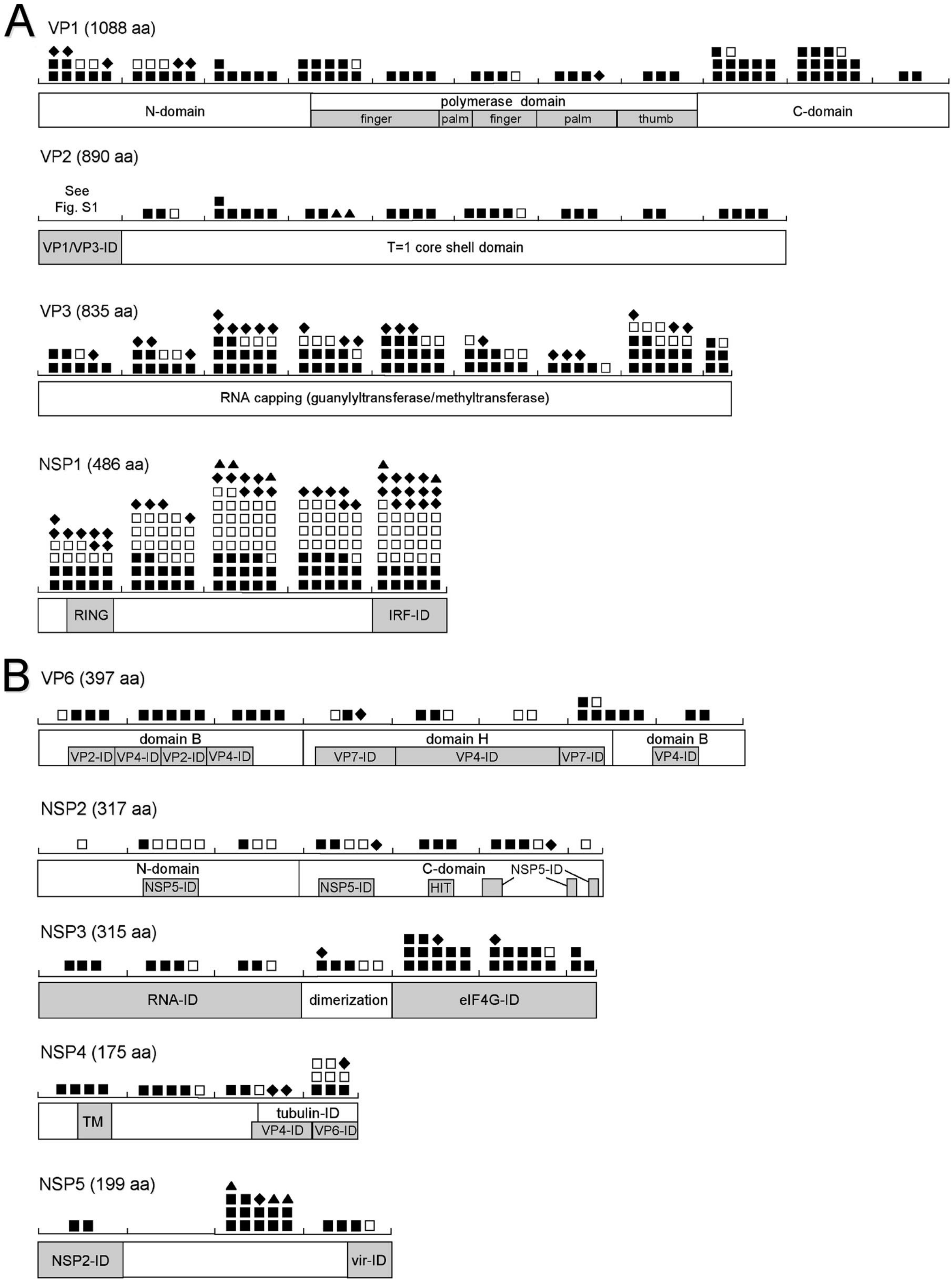


FIG. 3. Locations of genotype-specific changes in rotavirus proteins. Amino acid alignments of proteins encoded by HRVs with fully sequenced genomes were produced using MacVector 8.1.2 (see Fig. S1 in the supplemental material). From these alignments, residues that distinguished each of the three genotypes were identified. Such genotype-specific changes are indicated using symbols. Residues that are conserved among but differ between genotypes 1 and 2 are shown as black squares. Residues unique to genotype 3 proteins and not seen in genotype 1 or 2 are indicated as white squares. Residues that showed specific changes among genotypes 1, 2, and 3 are illustrated as black diamonds. Residues present in one genotype but not others (i.e., deletions) are indicated as black triangles. The amino acids identified in this analysis are presented in Table S1 in the supplemental material. For mapping purposes, the residues were then grouped into intervals of either (A) 100 amino acids (aa) (for VP1, VP2,

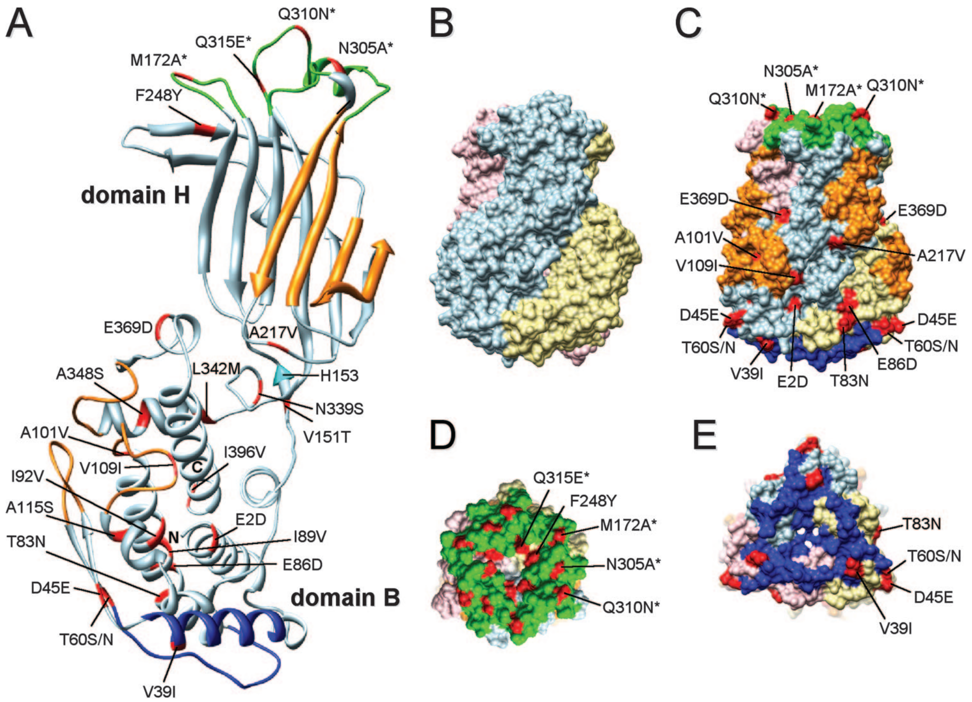


FIG. 4. Mapping genotype-specific differences on the atomic structure of VP6. (A) Ribbon diagram illustrating domain H (residues 151 to 331) and domain B (residues 1 to 150 and 331 to 397) of VP6 (bovine strain RF) (27). Regions of VP6 interfacing with VP2 (blue), VP4 (gold), and VP7 (green) are indicated, as are the locations of the N and C termini and the zinc-coordinating residue (cyan, H153). Amino acids differing between genotype 1 and 2 VP6 proteins are shown in red and are labeled. The genotype-specific residues in the VP7 interface region that define VP6 SG specificity are labeled with asterisks: SG-I (A172 and A305) and SG-II (M172, N305, Q310, and Q315) (17). (B) Surface side-view representation of the VP6 trimer, with monomers colored in light blue, pink, and yellow. (C) VP6 trimer as shown in panel B, with VP2, VP4, and VP7 interface regions and residues differing between genotypes 1 and 2 colored as described for panel A. (D and E) Top and bottom views, respectively, of the trimer shown in panel C.

specific epitopes (Fig. 4). In contrast, the site on VP6 where VP4 binds is quite conserved, and the β-roll of domain H does not show any differences between genotypes 1 and 2. Although genes encoding VP6, VP7, and VP4 of HRVs are suggested to reassort freely and independently in nature, some types of outer capsid proteins tend to appear together more frequently. For instance, genotype 2 VP6 proteins tend to associate most frequently with G2 VP7 and P[4] VP4 proteins. Preferential binding among these capsid proteins could influence (i) assembly of infectious virions, (ii) particle stability in nature, or (iii) efficiency of disassembly following entry into a host cell. Therefore, interactions between the outermost capsid proteins might be subjected to intrinsic selection pressures related to favorable interactions as well.

Interactions between VP6 and NSP4. In addition to binding to rotavirus structural proteins, VP6 engages the transmembrane glycoprotein (NSP4) during virion assembly. This interaction recruits newly formed DLPs into the endoplasmic reticulum, where, through a budding process, the outer capsid shell of the virion is formed (12). A previous study has provided evidence of a genetic linkage between VP6 and NSP4 (21). Specifically, HRVs exhibiting VP6 SG-II specificities (genotype 1) have proteins classified as NSP4B (genotype 1), whereas those bearing VP6 SG-I epitopes (genotype 2) encode NSP4A or NSP4C (genotype 2 or 3, respectively) (21). The results presented in this report are in agreement with this finding, as only three HRVs (N26-02, RV176-00, and RV161-00) for which sequences are available have VP6 and NSP4 proteins belonging to different genotypes (Fig. 1). Furthermore, while NSP4 contains very few genotype-specific amino acid differences, the changes cluster in the C terminus of the protein (Fig. 3). This observation is intriguing, as the C-terminal region of NSP4 is involved in interactions with VP6 (1). Thus, the genotype-specific amino acid variations may be connected to the genetic linkage noted for the VP6 and NSP4 proteins. Unfortunately, no information on the location of the NSP4-binding site on the VP6 trimer exists, but we expect the distal portion of the trimer to be the most likely site (Fig. 4).

VP3, and NSP1) or (B) 50 aa (for VP6, NSP2, NSP3, NSP4, and NSP5), spanning the length of the protein. The number of residues in each classification group per interval is plotted above the protein. Locations of key functional and structural domains are approximated. The genotype-specific changes for the VP2 N terminus are presented in Fig. S2 in the supplemental material. The NSP5 interactive domain (ID) on NSP2 was determined based on reference 22 and overlaps the RNA ID. IRF, interferon regulatory factor; HIT, histidine triad; TM, transmembrane domain; vir, viroplasm.

Nonetheless, NSP4 is a multifunctional protein that interacts not only with VP6 but also with VP4 and host cell membranes and proteins (e.g., tubulin) during viral replication. Aside from promoting rotavirus maturation, NSP4 functions as a viral enterotoxin that induces diarrhea through fluid loss triggered by intracellular calcium mobilization and chloride secretion (12). Thus, the pressures influencing the genetic diversity of NSP4 are not merely limited to viral protein interactions.

NSP2-NSP5 interactions and analysis of the NSP6 ORF. During replication inside a host cell, the nonstructural proteins NSP2 and NSP5 interact with each other to mediate the formation of viroplasms, which serve as sites of viral RNA synthesis and early virion morphogenesis (47). The RNA-binding protein NSP2 exists as a doughnut-shaped octamer that is formed by the interaction of two tetramers stacked tail to tail (22). Extending diagonally across the NSP2 octamer are four highly basic grooves; these represent the sites of interaction with NSP5 and single-stranded RNA (22). In addition to binding NSP5 and RNA, NSP2 is suggested to interact with one or more core proteins (e.g., VP1, VP2); yet, the binding sites for these proteins on NSP2 have not been determined (47). NSP2 and NSP5 are remarkably conserved among genotypes 1, 2, and 3, and nearly all of the genotype-specific changes for both of these proteins lie just outside of known interaction sites (Fig. 3). This observation is interesting in light of the fact that the NSP2- and NSP5-encoding genes seem to reassort more freely in the HRVs than do genes encoding the inner capsid proteins (Fig. 1). The conserved nature of the NSP2 and NSP5 interaction sites might allow these two proteins to function in a genotype-independent manner. In other words, NSP2 and NSP5 proteins from different genotypes (or even from different species) can functionally substitute for each other quite easily. For the phosphoprotein NSP5, the majority of the genotype-specific changes were noted to occur in the C-terminal half of the protein; the importance of this region is not known, but it might represent the binding site for cellular protein kinases (Fig. 3).

We should point out that, for many HRVs, the NSP5-coding gene (gene 11) has an alternative ORF that encodes a putative nonstructural protein (NSP6) (Fig. 1; also see Fig. S3 in the supplemental material). The existence of NSP6 in cells infected with HRVs has not been fully demonstrated. However, in studies using the simian strain SA11, NSP6 was detected using specific antisera and was shown to colocalize with NSP2 and NSP5 in viroplasms (33). The lengths of the NSP6 ORF sequences vary quite dramatically for the 25 HRVs analyzed in this study; yet, such variation has been noted for many animal rotavirus strains (33). Five HRVs (D, P, WI61, AU-1, and N26-02) show NSP6 ORFs that are truncated compared with that of SA11, while two of them (Se584 and A64) show extended NSP6 ORFs (Fig. 1; also see Fig. S3 in the supplemental material). We found no correlation between NSP6 ORF length and genotype/genogroup specificity. Future studies are needed to determine whether these HRVs truly express the proposed NSP6 proteins.

NSP1 and NSP3 interactions with host cells. The rotavirus nonstructural proteins NSP1 and NSP3 are unique in that they interact extensively with host cell factors but are not known to bind to any other viral proteins and are nonessential for replication in cell culture. As such, one might expect that NSP1 and NSP3 are subject to extrinsic selection pressures and should show patterns of evolution that differ from those of the other viral proteins. Our sequencing analysis shows that HRV NSP1 and NSP3 proteins can still be assigned into genotype 1, 2, or 3 but that the genes encoding these proteins tend to reassort more freely than those encoding the inner capsid proteins (Fig. 1). Additionally, the percent identity between genotypes 1 and 2 for these proteins is lower than that for any of the others, being 68.9% and 82.0% for NSP1 and NSP3, respectively, suggesting that these proteins are quite divergent (Table 3). NSP1 plays an important role in antagonizing the host cell innate immune response by inducing the proteosome-mediated degradation of transcription factors involved in interferon production (interferon regulatory factors) (3). Extensive genotype-specific changes for NSP1 are seen throughout the entire protein but are even more exaggerated in the C-terminal half (Fig. 3). A region within the C terminus of NSP1 is thought to be important for binding to interferon regulatory factor proteins (15). In contrast, the putative RING domain of NSP1, which contains a motif recognized in many E3 ubiquitin ligases, is quite conserved (Fig. 3). NSP3 is involved in abrogating host cell protein synthesis during infection and is known to interact with viral RNA and components of the cellular translation machinery (10, 40, 41). NSP3 is much more conserved than NSP1 but has genotype-specific changes clustering in its C terminus, a region important for interactions with eIF4G (eukaryotic initiation factor 4G) (Fig. 3). The reason for genotype-specific variation in NSP1 and NSP3 is not known, but it might reveal that these proteins participate in as-yet-unidentified interactions with other viral proteins.

In summary, the results presented in this report provide detailed information regarding the full-genome sequences of 10 commonly used HRV strains (D, DS-1, P, ST3, IAL28, Se584, 69M, WI61, A64, and L26). By utilizing and expanding upon the classification system of Matthijnssens et al. (30), we present evidence for the existence of preferred HRV gene constellations. Several previous reports have provided data to suggest that not all HRV genes reassort randomly; the existence of HRV genogroups is the strongest evidence in favor of this argument (9, 21, 34, 35, 49). Still, this study is the first to combine sequence analyses with structural and functional data to (i) gain insight into the normal variation within and between HRV genogroup proteins (i.e., genotypes) and (ii) provide a molecular explanation for why preferred gene constellations are maintained. Our results are consistent with the idea that genetic linkages are, at least in part, a result of intrinsic selection pressures on viral proteins. Nonetheless, we cannot rule out the possibility that interactions among RNA segments during gene assortment and packaging influence the constellations. It is important to note that intragenogroup reassortant viruses that break seemingly tight genetic linkages (for example, between VP2 and VP6) can be created in the laboratory, demonstrating that the preferences for intragenotypic protein interactions are subtle (data not shown). However, we predict that such reassortants, albeit viable, have decreased evolutionary fitness compared to that of parental strains and would be selected against in the context of infection in the human population. Without a doubt, further epidemiologic studies are required to examine if and how identified genotype-specific differences correlate with biological functions. We expect the results presented here to provide a scientific platform for such studies that will reveal the true dynamics of group A rotavirus evolution and diversity.

ACKNOWLEDGMENTS

We express our appreciation to Taka Hoshino for generously providing the HRVs used in this study and to Al Kapikian and Jelle Matthijnssens for insightful comments and critical readings of the manuscript.

This work was supported by the Intramural Research Program of the NIH, National Institute of Allergy and Infectious Diseases. E.M.H. was also supported by the Oak Ridge Associated Universities' Research Associates/Specialists Program at the NIH.

REFERENCES

1. **Ball, J. M., D. M. Mitchell, T. F. Gibbons, and R. D. Parr.** 2005. Rotavirus NSP4: a multifunctional viral enterotoxin. Viral Immunol. **18:**27–40.
2. **Ballard, A., M. A. McCrae, and U. Desselberger.** 1992. Nucleotide sequences of normal and rearranged RNA segments 10 of human rotaviruses. J. Gen. Virol. **73:**633–638.
3. **Barro, M., and J. T. Patton.** 2005. Rotavirus nonstructural protein 1 subverts innate immune response by inducing degradation of IFN regulatory factor 3. Proc. Natl. Acad. Sci. USA **102:**4114–4119.
4. **Charpilienne, A., J. Lepault, F. Rey, and J. Cohen.** 2002. Identification of rotavirus VP6 residues located at the interface with VP2 that are essential for capsid assembly and transcriptase activity. J. Virol. **76:**7822–7831.
5. **Chen, Y., Y. Wen, X. Liu, X. Xiong, Z. Cao, Q. Zhao, Y. Yu, X. Yin, C. Li, and Y. Fan.** 2008. Full genomic analysis of human rotavirus strain TB-Chen isolated in China. Virology **375:**361–373.
6. **Chrystie, I. L., B. Totterdell, M. J. Baker, J. W. Scopes, and J. E. Banatvala.** 1975. Letter: rotavirus infections in a maternity unit. Lancet **ii:**79.
7. **Clark, H. F., Y. Hoshino, L. M. Bell, J. Groff, G. Hess, P. Bachman, and P. A. Offit.** 1987. Rotavirus isolate WI61 representing a presumptive new human serotype. J. Clin. Microbiol. **25:**1757–1762.
8. Reference deleted.
9. **Cunliffe, N. A., J. S. Bresee, J. R. Gentsch, R. I. Glass, and C. A. Hart.** 2002. The expanding diversity of rotaviruses. Lancet **359:**640–642.
10. **Deo, R. C., C. M. Groft, K. R. Rajashankar, and S. K. Burley.** 2002. Recognition of the rotavirus mRNA 3′ consensus by an asymmetric NSP3 homodimer. Cell **108:**71–81.
11. **Desselberger, U., M. Iturriza-Gomara, and J. J. Gray.** 2001. Rotavirus epidemiology and surveillance. Novartis Found. Symp. **238:**125–152.
12. **Estes, M. K.** 2001. Rotaviruses and their replication, p. 1747–1785. *In* D. M. Knipe, P. M. Howley, D. E. Griffin, R. A. Lamb, M. A. Martin, B. Roizman, and S. E. Straus (ed.), Fields virology, 4th ed. Lippincott Williams & Wilkins, Philadelphia, PA.
13. **Gentsch, J. R., A. R. Laird, B. Bielfelt, D. D. Griffin, K. Banyai, M. Ramachandran, V. Jain, N. A. Cunliffe, O. Nakagomi, C. D. Kirkwood, T. K. Fischer, U. D. Parashar, J. S. Bresee, B. Jiang, and R. I. Glass.** 2005. Serotype diversity and reassortment between human and animal rotavirus strains: implications for rotavirus vaccine programs. J. Infect. Dis. **192:**S146–S159.
14. **Gouvea, V., and M. Brantly.** 1995. Is rotavirus a population of reassortants? Trends Microbiol. **3:**159–162.
15. **Graff, J. W., D. N. Mitzel, C. M. Weisend, M. L. Flenniken, and M. E. Hardy.** 2002. Interferon regulatory factor 3 is a cellular partner of rotavirus NSP1. J. Virol. **76:**9545–9550.
16. **Gray, J., and U. Desselberger.** 2000. Rotaviruses: methods and protocols. Humana Press, Totowa, NJ.
17. **Greig, S. L., J. A. Berriman, J. A. O'Brien, J. A. Taylor, A. R. Bellamy, M. J. Yeager, and A. K. Mitra.** 2006. Structural determinants of rotavirus subgroup specificity mapped by cryo-electron microscopy. J. Mol. Biol. **356:**209–221.
18. **Griffin, D. D., C. D. Kirkwood, U. D. Parashar, P. A. Woods, J. S. Bresee, R. I. Glass, J. R. Gentsch, and the National Rotavirus Strain Surveillance System Collaborating Laboratories.** 2000. Surveillance of rotavirus strains in the United States: identification of unusual strains. J. Clin. Microbiol. **38:**2784–2787.
19. **Hasegawa, A., S. Inouye, S. Matsuno, K. Yamaoka, R. Eko, and W. Suharyono.** 1984. Isolation of human rotaviruses with a distinct RNA electrophoretic pattern from Indonesia. Microbiol. Immunol. **28:**719–722.
20. **Hundley, F., M. McIntyre, B. Clark, G. Beards, D. Wood, I. Chrystie, and U. Desselberger.** 1987. Heterogeneity of genome rearrangements in rotaviruses isolated from a chronically infected immunodeficient child. J. Virol. **61:**3365–3372.
21. **Iturriza-Gòmara, M., E. Anderton, G. Kang, C. Gallimore, W. Phillips, U. Desselberger, and J. Gray.** 2003. Evidence for genetic linkage between the gene segments encoding NSP4 and VP6 proteins in common and reassortant human rotavirus strains. J. Clin. Microbiol. **41:**3566–3573.
22. **Jayaram, H., Z. Taraporewala, J. T. Patton, and B. V. Prasad.** 2002. Rotavirus protein involved in genome replication and packaging exhibits a HIT-like fold. Nature **417:**311–315.
23. **Kapikian, A. Z., Y. Hoshino, and R. M. Chanock.** 2001. Rotaviruses, p. 1787–1833. *In* D. M. Knipe, P. M. Howley, D. E. Griffin, R. A. Lamb, M. A. Martin, B. Roizman, and S. E. Straus (ed.), Fields virology, 4th ed. Lippincott Williams & Wilkins, Philadelphia, PA.
24. **Kapikian, A. Z., H. W. Kim, R. G. Wyatt, W. J. Rodriguez, S. Ross, W. L. Cline, R. H. Parrott, and R. M. Chanock.** 1974. Reoviruslike agent in stools: association with infantile diarrhea and development of serologic tests. Science **185:**1049–1053.
25. **Kapikian, A. Z., L. Simonsen, T. Vesikari, Y. Hoshino, D. M. Morens, R. M. Chanock, J. R. La Montagne, and B. R. Murphy.** 2005. A hexavalent human rotavirus-bovine rotavirus (UK) reassortant vaccine designed for use in developing countries and delivered in a schedule with the potential to eliminate the risk of intussusception. J. Infect. Dis. **192:**S22–S29.
26. **Lin, S. L., and P. Tian.** 2003. Detailed computational analysis of a comprehensive set of group A rotavirus NSP4 proteins. Virus Genes **26:**271–282.

26a. **Lu, X., S. M. McDonald, M. A. Tortorici, Y. J. Tao, R. Vasquez-Del Carpio, M. L. Nibert, J. T. Patton, and S. C. Harrison.** Mechanism for coordinated RNA packaging and genome replication by rotavirus polymerase VP1. Structure, in press.

27. **Mathieu, M., I. Petitpas, J. Navaza, J. Lepault, E. Kohli, P. Pothier, B. V. Prasad, J. Cohen, and F. A. Rey.** 2001. Atomic structure of the major capsid protein of rotavirus: implications for the architecture of the virion. EMBO J. **20:**1485–1497.
28. **Matsui, S. M., E. R. Mackow, S. Matsuno, P. S. Paul, and H. B. Greenberg.** 1990. Sequence analysis of gene 11 equivalents from "short" and "super short" strains of rotavirus. J. Virol. **64:**120–124.
29. **Matsuno, S., A. Hasegawa, A. Mukoyama, and S. Inouye.** 1985. A candidate for a new serotype of human rotavirus. J. Virol. **54:**623–624.
30. **Matthijnssens, J., M. Ciarlet, E. Heiman, I. Arijs, T. Delbeke, S. M. McDonald, E. A. Palombo, M. Iturriza-Gòmara, P. Maes, J. T. Patton, M. Rahman, and M. Van Ranst.** 2008. Full genome-based classification of rotaviruses reveals a common origin between human Wa-like and porcine rotavirus strains and human DS-1-like and bovine rotavirus strains. J. Virol. **82:**3204–3219.
31. **Matthijnssens, J., M. Rahman, X. Yang, T. Delbeke, I. Arijs, J. P. Kabue, J. J. Muyembe, and M. Van Ranst.** 2006. G8 rotavirus strains isolated in the Democratic Republic of Congo belong to the DS-1-like genogroup. J. Clin. Microbiol. **44:**1801–1809.
32. **McDonald, S. M., and J. T. Patton.** 2008. Molecular characterization of a subgroup specificity associated with the rotavirus inner capsid protein VP2. J. Virol. **82:**2752–2764.
33. **Mohan, K. V., R. I. Glass, and C. D. Atreya.** 2006. Comparative molecular characterization of gene segment 11-derived NSP6 from lamb rotavirus LLR strain used as a human vaccine in China. Biologicals **34:**265–272.
34. **Nakagomi, O., and T. Nakagomi.** 2002. Genomic relationships among rotaviruses recovered from various animal species as revealed by RNA-RNA hybridization assays. Res. Vet. Sci. **73:**207–214.
35. **Nakagomi, O., T. Nakagomi, K. Akatani, and N. Ikegami.** 1989. Identification of rotavirus genogroups by RNA-RNA hybridization. Mol. Cell. Probes **3:**251–261.
36. **Nakagomi, O., T. Nakagomi, Y. Hoshino, J. Flores, and A. Z. Kapikian.** 1987. Genetic analysis of a human rotavirus that belongs to subgroup I but has an RNA pattern typical of subgroup II human rotaviruses. J. Clin. Microbiol. **25:**1159–1164.
37. **Parashar, U. D., E. G. Hummelman, J. S. Bresee, M. A. Miller, and R. I. Glass.** 2003. Global illness and deaths caused by rotavirus disease in children. Emerg. Infect. Dis. **9:**565–572.
38. **Patton, J. T., M. T. Jones, A. N. Kalbach, Y. W. He, and J. Xiaobo.** 1997. Rotavirus RNA polymerase requires the core shell protein to synthesize the double-stranded RNA genome. J. Virol. **71:**9618–9626.
39. **Pesavento, J. B., S. E. Crawford, M. K. Estes, and B. V. Prasad.** 2006. Rotavirus proteins: structure and assembly. Curr. Top. Microbiol. Immunol. **309:**189–219.
40. **Piron, M., T. Delaunay, J. Grosclaude, and D. Poncet.** 1999. Identification of the RNA-binding, dimerization, and eIF4GI-binding domains of rotavirus nonstructural protein NSP3. J. Virol. **73:**5411–5421.
41. **Piron, M., P. Vende, J. Cohen, and D. Poncet.** 1998. Rotavirus RNA-binding protein NSP3 interacts with eIF4GI and evicts the poly(A) binding protein from eIF4F. EMBO J. **17:**5811–5821.
42. **Pongsuwanna, Y., R. Guntapong, M. Chiwakul, R. Tacharoenmuang, N. Onvimala, M. Wakuda, N. Kobayashi, and K. Taniguchi.** 2002. Detection of a human rotavirus with G12 and P[9] specificity in Thailand. J. Clin. Microbiol. **40:**1390–1394.
43. **Rahman, M., J. Matthijnssens, X. Yang, T. Delbeke, I. Arijs, K. Taniguchi, M. Iturriza-Gòmara, N. Iftekharuddin, T. Azim, and M. Van Ranst.** 2007. Evolutionary history and global spread of the emerging G12 human rotaviruses. J. Virol. **81:**2382–2390.
44. **Ramachandran, M., C. D. Kirkwood, L. Unicomb, N. A. Cunliffe, R. L. Ward, M. K. Bhan, H. F. Clark, R. I. Glass, and J. R. Gentsch.** 2000. Molecular

characterization of serotype G9 rotavirus strains from a global collection. Virology **278:**436–444.
45. **Rodriguez, W. J., H. W. Kim, C. D. Brandt, R. H. Yolken, J. O. Arrobio, A. Z. Kapikian, R. M. Chanock, and R. H. Parrott.** 1978. Sequential enteric illnesses associated with different rotavirus serotypes. Lancet **ii:**37.
46. **Taniguchi, K., T. Urasawa, N. Kobayashi, M. Gorziglia, and S. Urasawa.** 1990. Nucleotide sequence of VP4 and VP7 genes of human rotaviruses with subgroup I specificity and long RNA pattern: implication for new G serotype specificity. J. Virol. **64:**5640–5644.
47. **Taraporewala, Z. F., and J. T. Patton.** 2004. Nonstructural proteins involved in genome packaging and replication of rotaviruses and other members of the Reoviridae. Virus Res. **101:**57–66.
48. **Timenetsky Mdo, C., V. Gouvea, N. Santos, R. C. Carmona, and Y. Hoshino.** 1997. A novel human rotavirus serotype with dual G5–G11 specificity. J. Gen. Virol. **78:**1373–1378.
49. **Ward, R. L., O. Nakagomi, D. R. Knowlton, M. M. McNeal, T. Nakagomi, J. D. Clemens, D. A. Sack, and G. M. Schiff.** 1990. Evidence for natural reassortants of human rotaviruses belonging to different genogroups. J. Virol. **64:**3219–3225.
50. **Wyatt, R. G., H. D. James, Jr., A. L. Pittman, Y. Hoshino, H. B. Greenberg, A. R. Kalica, J. Flores, and A. Z. Kapikian.** 1983. Direct isolation in cell culture of human rotaviruses and their characterization into four serotypes. J. Clin. Microbiol. **18:**310–317.
51. **Wyatt, R. G., W. D. James, E. H. Bohl, K. W. Theil, L. J. Saif, A. R. Kalica, H. B. Greenberg, A. Z. Kapikian, and R. M. Chanock.** 1980. Human rotavirus type 2: cultivation in vitro. Science **207:**189–191.
52. **Zeng, C. Q., M. K. Estes, A. Charpilienne, and J. Cohen.** 1998. The N terminus of rotavirus VP2 is necessary for encapsidation of VP1 and VP3. J. Virol. **72:**201–208.

View publication stats

Središnja medicinska knjižnica

Habek M., Borovečki F., Brinar V. V. (2010) ***Genomics in multiple sclerosis.*** **Clinical Neurology and Neurosurgery, [Epub ahead of print]. ISSN 0303-8467**

http://www.elsevier.com/locate/issn/03038467

http://www.sciencedirect.com/science/journal/03038467

http://dx.doi.org/10.1016/j.clineuro.2010.03.028

http://medlib.mef.hr/813

University of Zagreb Medical School Repository

http://medlib.mef.hr/

Genomics in multiple sclerosis

Mario Habek, MD[1,2], Fran Borovečki, MD, PhD[2,3], Vesna V Brinar, MD, PhD,[1,2]

School of Medicine, University of Zagreb and University Hospital Center Zagreb, [1]Department of Neurology, [2]Refferal Center fo Demyelinating Diseases of the Central Nervous System, [3]Department for Functional Genomics, Center for Translational and Clinical Research, Zagreb, Croatia

Corresponding author:

Mario Habek, MD
University Department of Neurology
Zagreb School of Medicine and University Hospital Center
Kišpatićeva 12
HR-10000 Zagreb
Croatia
Phone: +38598883323; Fax: +38512421891; e-mail: mhabek@mef.hr

Word count: 2970

Conflict of interest statement: There is no conflict of interest.

Abstract

Multiple sclerosis (MS) is chronic, inflammatory disease of the central nervous system that mainly affects young adults and is characterized with dissemination of demyelinating lesions in time and space. It is well known that MS is very heterogeneous disease, so biomarkers that would reliably determine disease course, outcome or treatment response in early stages of the disease (preferentially clinically isolated syndrome) are desperately needed. Genome-wide expression analysis represents the profile or imprint of all genes in a certain tissue or cell population in certain time point. Therefore, as human genome is entirely known, it is possible to analyze any given human gene in any given context. This review will discuss results and possible applications of gnome-wide expression studies in brain tissue and blood samples of MS patients.

Key words: Multiple sclerosis, biomarkers, genomics, microarrays

Introduction

Multiple sclerosis (MS) is chronic, inflammatory disease of the central nervous system (CNS) that mainly affects young adults and is characterized with demyelination. Although MS is pathologically and clinically heterogeneous disease, in 85% of patients it starts with acute or subacute episode of neurological dysfunction attributable to one or more foci of demyelination, which is called clinically isolated syndrome (CIS). In 21% of patients, CIS manifests as optic neuritis, 46% myelopathy, 10% with brainstem symptoms and in 23% with multifocal symptoms.[1] One of the main characteristics of MS is high heterogeneity in final outcome between patients, with benign or even asymptomatic disease on one end and highly aggressive, malignant disease on the other end of the spectrum. Therefore biomarkers that would reliably determine disease course, outcome or treatment response in early stages of the disease (preferentially CIS) are desperately needed.

Biomarkers

Biomarkers are very important indicators of normal or abnormal biological processes. By definition, biological markers or biomarkers are characteristics that can be objectively measured, and serve as indicators of normal biological process, pathological process or pharmacological response to therapeutic intervention.[2] Potential implications for biomarkers are numerable, they can be used in disease diagnosis and monitoring and determining early efficiency of treatment. As well, they are invaluable in early disease detection, staging and prognosis.

MS is heterogeneous disease in rate of progression, clinical symptoms, specificity of immune response and pathology of MS lesions, so it is to be expected that certain biomarker will reflect just one of many pathogenetic mechanisms involved.[3] Therefore patient stratification will be possible only with defined group of biomarkers so relative contribution of each of different pathogenetic mechanisms can be determined.[4] It is also important that every biomarker is validated in different, independent cohorts in prospective multicentric studies.

Validation criteria for each biomarker are defined according to the purpose for which the biomarker in question is created and should be standardized in different cohorts with following goals:[4]

1) Clinical relevance: biomarker has ability to follow changes in the pathological process and/or therapeutic intervention in relatively short period of time
2) Sensitivity: ability that the biomarker can be measured with precisely enough and that change of biomarker reflects the change of clinical endpoint
3) Specificity: ability of biomarker to identify persons with certain disease or response to certain therapeutic intervention
4) Probability of falsely positive and falsely negative cases: defined as situation in which change of the biomarker does not reflect change of the certain clinical endpoint
5) Accuracy, precision, reproducibility and variability of the laboratory test which measures the biomarker

Currently there is no MS biomarker that fulfills all of the above mentioned criteria.[5]

Genomics

The development of genomics has for the first time made possible to overcome many of above mentioned problems in development of MS biomarkers. With this method it is possible to determine the profile or imprint of all genes in certain tissue or cell population in certain time point. As human genome is entirely known, it is possible to analyze any given human gene in any given context, so it is possible to understand complex molecular interactions, discover biomarkers that correlate well with clinical signs and discover new therapeutic possibilities.[6] Genomic analysis is performed with microarrays that contain known oligonucleotides of very high density, attached to surface in a specific order of very high density. DNA microarrays can be divided in two groups according to DNA type: oligo microarrays contain synthesized oligonucleotids, while cDNA microarrays contain complementary DNA molecule cloned or amplified with PCR.[7,8] The main objection of this method is that it is not hypothesis driven. However, because genetically MS is very complex disease and there is great heterogeneity in clinical picture, genome-wide expression studies have clear advantage over conventional studies, with possibility of forming new hypotheses.[6] During the interpretation of data obtained with these studies, researcher should always bear in mind intra- and interindividual variations between patients[9]. Furthermore, there always exists a question of technical variability, namely use of different microarrays and manufacturer specific protocols, which may influence the final results of expression profiling experiments. Recent studies, such as Microarray Quality Control project have provided some reassurance about the reproducibility of contemporary microarray platforms, showing an average 89% overlap in expression profiles generated between sites using the same microarray platforms and 74% overlap across platforms from different manufacturers.[10] Alternatively, using two or more microarray platforms in analysis of the same samples and selecting the most reproducibly differentially expressed genes as biomarkers, could provide a way to reduce the influence of inter-platform technical variability on the process of biomarker selection. Finally, proper selection of statistical methods used is crucial, as it can also be a source of bias in the procedure of new biomarker selection.

Recently guidelines have been set out for improvement of reliability of microarray results:[6]

1. Microarray results should be confirmed with alternative methods (real time PCR). However shortage of confirmatory tests does not reduce the importance of microarray results if the microarray experiment is of sufficient quality.
2. Microarray results should always be replicated in an independent cohort of patients. As an alternative one may use biostatistical methods.
3. All samples should be collected at the same time of the day, and time between sample collection and processing should be standardized.
4. If one is doing blood samples analysis, distribution of main mononuclear cell lineages should be analyzed.
5. Relapses, infections and drugs should not interfere with the results.
6. All samples should be collected by standardized method.
7. Analysis should include and dose genes which are near the cut-ff values.
8. Microarrays provide us with great amount of data, so every researcher should make a repository of all available data so the rest of scientific community can gain access to the results. Repositories like this should be in concordance with MIAME (Minimum Information About a Microarray Experiment) guidelines.[11]

What have we learned from genome-wide expression studies in brain tissue of MS patients?

The first study using gene microarrays on brain tissue of MS patients was published in 1999.[12] This study investigated expression profiles differences of more than 5000 genes between normal brain tissue and acute MS plaques and found expression differences in 62 genes, among them Duffy chemokine receptor, interferon regulatory factor-2 and TNF alpha receptor2. Following this study, many investigators used gene microarrays on brain tissue of MS patients with aim of discovery of new biomarkers[13,14], new pathophysiological mechanisms or new therapeutic targets.[9,15-22]

Most of these studies are performed on postmortem tissue, because biopsies are very rarely performed in MS patients. As degradation of RNA happens relatively quickly after death, it is necessary to process brain tissue very quickly in order to obtain RNA of high enough quality for DNA microarrays. Despite this RNA can be isolated from postmortem samples up to 20 hours after death, although short time does not guaranty RNA of high quality.[23] It should be emphasized that postmortem tissue, despite all disadvantages, is a very important source for gene microarrays, especially when investigating new pathogenetic mechanisms and discovery of new therapeutic targets.[23] It is always necessary to carefully characterize the tissue from which RNA will be isolated, immunohistochemistry with markers for demyelination, infiltration and gliosis are all required for good interpretation of results. All these tasks are almost impossible to fulfill, but they should be taken into account when analyzing the data.

In order to identify genes which are crucial in MS pathogenesis, Whitney and coworkers have compared gene expression profiles in MS lesions and brains of EAE mice with normal white matter.[12] Altogether 2798 genes were compared and one of the most important findings was that 5-lipooxigenase, a key enzyme in the biosynthesis of proinlammatory leukotrienes, is overexpressed in MS lesions and brains of EAE mice. Although this finding is not specific to

MS, it emphasizes the importance of proinflammatory activity in demyelinating process and suggests the possible role of antiiflammatory therapy in MS.

Tajouri and coworkers have compared RNA expression profiles of chronic active and acute MS lesions between themselves and in comparison with normal white matter.[15] These authors have identified 139 differentially expressed genes between MS lesion and normal tissue, 69 of them showed same direction of expression in both chronic active and acute MS lesions, while 70 genes showed different direction of expression between chronic active and acute MS lesions. This study revealed genes with already known role in MS pathogenesis like myelin basic protein, glutathione-S-transferase M1 and different growth factors. As both, chronic and acute lesion, had similar expression profiles, it has been suggested that quantitative rather than qualitative differences in gene expression, define progression from acute to chronic active MS lesion.

Further studies have confirmed differences in gene expression between chronic active and inactive MS lesions, in their edges as well as in the central parts.[16] The most differentially expressed genes were genes implicated in inflammatory response, apoptosis related and stress-induced genes. Major differences in gene expression occurred between the lesion margin and lesion centre in active lesions (57 and 69 genes differentially expressed, respectively), whereas the margins and centres of silent lesions showed markedly reduced heterogeneity (only 11 and two genes differentially expressed, respectively).[17]

Microarray analysis of MS lesions obtained at autopsy revealed increased transcripts of genes encoding inflammatory cytokines, particularly interleukin-6 and -17, interferon-gamma and associated downstream pathways. Comparison of acute lesions with inflammation versus 'silent' lesions without inflammation also revealed differentially transcribed genes.[18] Some of these differentially expressed genes were chosen as therapeutic targets in EAE model of MS, and these results have confirmed the importance of microarray analysis research.

Microarrays on autopsy samples of patients with secondary progressive MS provided molecular evidence of a continuum of dysfunctional homeostasis and inflammatory changes between lesions and normal appearing white matter (NAWM), and supported the concept of MS pathogenesis being a generalised process that involves the entire CNS.[19] When comparing gene expression in NAWM of postmortem brains of MS patients, there is increased expression of genes involved in maintenance of cellular homeostasis, and in neural protective mechanisms known to be induced upon ischemic preconditioning.[20] When comparing expression levels of 33,000 characterized genes in postmortem motor cortex from MS brains, compared with controls, 488 transcripts were found to be decreased and 67 increased. Twenty-six nuclear-encoded mitochondrial genes and the functional activities of mitochondrial respiratory chain complexes I and III were decreased in the MS motor cortex, which was specific for neurons. In addition, pre-synaptic and postsynaptic components of GABAergic neurotransmission and the density of inhibitory interneuron processes also were decreased in the MS cortex.[21] These data supports a mechanism whereby reduced ATP production in demyelinated segments of upper motor neuron axons contributes to progressive neurological disability in MS patients.

One of the advantages of microarray technology is that specific group of genes related to a specific pathway or process can be analyzed separately. In this context, a focused endothelial cell biology microarray, capable of detecting changes in expression of 113 blood/brain barrier-specific genes showed 52 differentially expressed genes in MS lesions compared to NAWM and healthy controls.[22]

What have we learned from gene expression studies on human blood samples in MS?

One of the most important factor that one should take into account when interpreting results of gene expression studies on human blood samples is great interindividual and time variation in gene expression in healthy subjects.[24,25] These variations are dependent on cell composition of the blood, sex, age and time of the day when samples are taken. The partly intrinsic variations can be caused by genotype differences, epigenetic phenomena or environmental and nutritional factors. That is why as many as possible variables should be controlled during processing of samples for microrarray studies.

On the other side, gene expression studies on human blood samples have many advantages. Blood is a very accessible sample to take so it is possible to perform analysis on larger number of patients than in studies using brain tissue. Results obtained with this method can give insight into drug efficacy much earlier, because blood is one of the first tissue with which the drug comes into touch with. However it should be kept in mind that blood also has many disadvantages, mainly because it reflects influence of other factors not pertinent to the disease process itself.[26]

There are two possible ways how to isolate RNA from blood samples. One is isolation of peripheral blood mononuclear cells (PBMC) from the whole blood by Ficoll gradient, and the other one is isolation of RNA from the whole blood with Paxgene tubes. Both procedures have their advantages and disadvantages.

Gene expression studies on human blood samples in MS have been used for diagnostic (differentiation of MS patients from healthy controls, differentiation different MS types, differentiation of treated MS patients from untreated MS patients), prognostic (differentiation of MS patients in relapse from patients in remission) and therapeutic (differentiation of treatment responders form treatment failures) purposes. One of the first studies aimed at discovering specific expression patterns in blood of patients demonstrated that cDNA microarray technology could be used to identify large-scale abnormal gene expression patterns in the peripheral blood of MS patients. This study identified 34 differentially expressed genes in patients with a relapse-remitting form of the disease when compared to healthy individuals. Significant increases were observed in molecules involved in T-cell and B-cell activation (LCK, CAMP responsive element modulator, IL-7 receptor), and degradation of extracellular matrix (MMP-19 or RASI-1), and significant decreases were observed in proteins that serve as chemokine receptors (STRL 22), are involved in apoptosis (DNA fragmentation factor-45) or in humoral immune responses (immunoglobulin heavy chain Gm marker).[27] Further to this study, Bomprezzi and collaborators identified a set of 53 genes differentially expressed in MS patients that can be used to predict the disease state in an independent test set.[13] As well the findings of this study supported the significance of autoreactive T cell activation as a primary pathophysiological event in MS. In an effort to develop a biomarker capable on not only successfully diagnosing patients with MS, but also differentiating between those experiencing a relapse from patients in remission, Achiron and coworkers identified 721 genes involved in activation of T-cells, epitope spreading and evasion of immune regulation in patients with acute MS relapse, when compared to MS patients in remission.[28] As well, this study broadened the number of differentiating genes between MS and healthy controles to 1,109 (589 overexpressed and 520 underexpressed) mainly involved in T-cell expansion and activation, inflammatory stimuli (cytokines and integrins), epitope spreading, and survival advantage leading to aberrant apoptosis.

Much larger number of studies tried to investigate changes in peripheral blood expression in response to immunomodulatory therapy. A study performed in patients receiving interferon beta therapy discovered that its action is not purely anti-inflammatory[29], and further studies identified 21 genes differentially expresses in blood of MS patients treated with interferon beta after 3 and 6 months of therapy. Out of the 21 differentially expressed genes, 9 possessed interferon responsive promoters. No significant change in expression of Th1 or Th2 related genes was observed.[30] These studies showed the complexity of expression changes in response to interferon therapy, and indicated the effect it exerts on cell migration, matrix degradation, proliferation, cell cycle control, differentiation, cell processing and presentation, apoptosis and cytokine and chemokine regulation. Trying to identify genes which could predict a favorable response to interferon therapy, a group of authors showed that interleukin 8 might be useful in predicting which patients will show a positive response.[31] These studies also showed that interferon beta induces expression of genes in a selective and time dependent manner, indicating its possible role in monitoring response to therapy.[32] This is especially true for genes involved in antiviral response, which are induced 1-4 hours following interferon beta administration. Similar expression pattern can be observed in genes involved in interferon beta signaling and lymphocyte activation.

Meta-analysis of expression profiling studies conducted using blood samples of patients with MS indicated 15 potential biomarkers which are expressed during the entire course of beta interferon treatment and which can serve as biomarkers of response to therapy. These include EIF2AK2, IFI6, IFI44, IFI44L, IFIH1, IFIT1, IFIT2, IFIT3, ISG15, MX1, OASL, RSAD2, SN, XAF1 and transcript represented by the Affymetrix probe 238704_at.[33] These biomarkers were all formerly identified as being indicative for IFNB activity.

More recently, it has been shown that after second day, one month, 12 months and 24 months of initiation of interferon therapy there are 42, 175, 103 and 108 differentially expressed genes, respectively. *MS4A1* (CD20), a known target of B-cell depletion therapy, was significantly downregulated after one month and *CMPK2*, *FCER1A*, and *FFAR2* appeared as hitherto unrecognized multiple sclerosis treatment-related differentially expressed genes that were consistently modulated over time.[34]

Another drug approved for RRMS treatment is glatiramer actetate (GA). Gene expression studies on human blood samples have shown that GA alters expression of 480 genes within 3 months of treatment; 262 genes were up-regulated, and 218 genes were down-regulated. The main convergent mechanisms of GA effects were related to antigen-activated apoptosis, inflammation, adhesion, and MHC class-I antigen presentation.[35]

Conclusion

So far, microarrays performed on brain tissue of MS patient provided us with important findings related to MS pathogenesis. Identification of endogenous protective mechanisms in MS brains can lead to development of new drugs which could slow the progression of disability.[23]

The expression signature of MS as detected in PBMC can be used in assessment and monitoring MS patients, discovery of new pathogenetic mechanisms and monitoring of response to treatment. However one has to bare in mind, that in a complex disease like MS both multiple interactions of different components of the immune system *in vivo*, and the complexity of the intracellular pathways must be considered in the interpretation of microarray experiments.[36]

In both cases, future studies with larger number of well defined patients will give us a better insight in molecular basis of heterogeneity of MS.

References

1. Confavreux C, Vukusic S, Moreau T, Adeleine P. Relapses and progression of disability in multiple sclerosis. N Engl J Med 2000;343:1430-8
2. Biomarkers Definitions Working Group. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clin Pharmacol Ther 2001;69:89-95
3. Weiner HL. Multiple sclerosis is an inflammatory T-cell-mediated autoimmune disease. Arch Neurol 2004;61:1613–5
4. Lutterotti A, Berger T, Reindl M. Biological markers for multiple sclerosis. Curr Med Chem 2007;14:1956-65
5. Bielekova B, Martin R. Development of biomarkers in multiple sclerosis. Brain 2004;127:1463-78.
6. Comabella M, Martin R. Genomics in multiple sclerosis--current state and future directions. J Neuroimmunol 2007;187:1-8
7. Duggan DJ, Bittner M, Chen Y, Meltzer P, Trent JM. Expression profiling using cDNA microarrays. Nat Genet 1999;21:10-4.
8. Barrett JC, Kawasaki ES. Microarrays: the use of oligonucleotides and cDNA for the analysis of gene expression. Drug Discov Today 2003;8:134-41
9. Whitney LW, Ludwin SK, McFarland HF, Biddison WE. Microarray analysis of gene expression in multiple sclerosis and EAE identifies 5-lipoxygenase as a component of inflammatory lesions. J Neuroimmunol 2001;121:40–48
10. MAQC Consortium. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nat Biotechnol 2006;24:1151-61

11. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FC, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson A, Sarkans U, Schulze-Kremer S, Stewart J, Taylor R, Vilo J, Vingron M. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet 2001;29:365-71
12. Whitney LW, Becker KG, Tresser NJ, Caballero-Ramos CI, Munson PJ, Prabhu VV, Trent JM, McFarland HF, Biddison WE. Analysis of gene expression in mutiple sclerosis lesions using cDNA microarrays. Ann Neurol 1999;46:425-8
13. Bomprezzi R, Ringnér M, Kim S, Bittner ML, Khan J, Chen Y, Elkahloun A, Yu A, Bielekova B, Meltzer PS, Martin R, McFarland HF, Trent JM. Gene expression profile in multiple sclerosis patients and healthy controls: identifying pathways relevant to disease. Hum Mol Genet 2003;12:2191-9
14. Mandel M, Gurevich M, Pauzner R, Kaminski N, Achiron A. Autoimmunity gene expression portrait: specific signature that intersects or differentiates between multiple sclerosis and systemic lupus erythematosus. Clin Exp Immunol 2004;138:164-70
15. Tajouri L, Mellick AS, Ashton KJ, Tannenberg AE, Nagra RM, Tourtellotte WW, Griffiths LR. Quantitative and qualitative changes in gene expression patterns characterize the activity of plaques in multiple sclerosis. Brain Res Mol Brain Res 2003;119:170-83
16. Mycko MP, Papoian R, Boschert U, Raine CS, Selmaj KW. Microarray gene expression profiling of chronic active and inactive lesions in multiple sclerosis. Clin Neurol Neurosurg 2004;106:223-9

17. Mycko MP, Papoian R, Boschert U, Raine CS, Selmaj KW. cDNA microarray analysis in multiple sclerosis lesions: detection of genes associated with disease activity. Brain 2003;126:1048-57
18. Lock C, Hermans G, Pedotti R, Brendolan A, Schadt E, Garren H, Langer-Gould A, Strober S, Cannella B, Allard J, Klonowski P, Austin A, Lad N, Kaminski N, Galli SJ, Oksenberg JR, Raine CS, Heller R, Steinman L. Gene-microarray analysis of multiple sclerosis lesions yields new targets validated in autoimmune encephalomyelitis. Nat Med 2002;8:500-8
19. Lindberg RL, De Groot CJ, Certa U, Ravid R, Hoffmann F, Kappos L, Leppert D. Multiple sclerosis as a generalized CNS disease--comparative microarray analysis of normal appearing white matter and lesions in secondary progressive MS. J Neuroimmunol 2004;152:154-67
20. Graumann U, Reynolds R, Steck AJ, Schaeren-Wiemers N. Molecular changes in normal appearing white matter in multiple sclerosis are characteristic of neuroprotective mechanisms against hypoxic insult. Brain Pathol 2003;13:554-73
21. Dutta R, McDonough J, Yin X, Peterson J, Chang A, Torres T, Gudz T, Macklin WB, Lewis DA, Fox RJ, Rudick R, Mirnics K, Trapp BD. Mitochondrial dysfunction as a cause of axonal degeneration in multiple sclerosis patients. Ann Neurol 2006;59:478-89
22. Cunnea P, McMahon J, O'Connell E, Mashayekhi K, Fitzgerald U, McQuaid S. Gene expression analysis of the microvascular compartment in multiple sclerosis using laser microdissected blood vessels. Acta Neuropathol 2009 [Epub ahead of print]
23. Kinter J, Zeis T, Schaeren-Wiemers N. RNA profiling of MS brain tissues. Int MS J 2008;15:51-8

24. Whitney AR, Diehn M, Popper SJ, Alizadeh AA, Boldrick JC, Relman DA, Brown PO. Individuality and variation in gene expression patterns in human blood. Proc Natl Acad Sci U S A 2003;100:1896-901.
25. Radich JP, Mao M, Stepaniants S, Biery M, Castle J, Ward T, Schimmack G, Kobayashi S, Carleton M, Lampe J, Linsley PS. Individual-specific variation of gene expression in peripheral blood leukocytes. Genomics 2004;83:980-8
26. Goertsches R, Zettl UK. MS therapy research applying genome-wide RNA profiling of peripheral blood. Int MS J 2007;14:98-107
27. Ramanathan M, Weinstock-Guttman B, Nguyen LT, Badgett D, Miller C, Patrick K, Brownscheidle C, Jacobs L. In vivo gene expression revealed by cDNA arrays: the pattern in relapsing-remitting multiple sclerosis patients compared with normal subjects. J Neuroimmunol 2001;116:213-9
28. Achiron A, Gurevich M, Friedman N, Kaminski N, Mandel M. Blood transcriptional signatures of multiple sclerosis: unique gene expression of disease activity. Ann Neurol 2004;55:410-7
29. Wandinger KP, Stürzebecher CS, Bielekova B, Detore G, Rosenwald A, Staudt LM, McFarland HF, Martin R. Complex immunomodulatory effects of interferon-beta in multiple sclerosis include the upregulation of T helper 1-associated marker genes. Ann Neurol 2001;50:349-57
30. Koike F, Satoh J, Miyake S, Yamamoto T, Kawai M, Kikuchi S, Nomura K, Yokoyama K, Ota K, Kanda T, Fukazawa T, Yamamura T. Microarray analysis identifies interferon beta-regulated genes in multiple sclerosis. J Neuroimmunol 2003;139:109-18

31. Stürzebecher S, Wandinger KP, Rosenwald A, Sathyamoorthy M, Tzou A, Mattar P, Frank JA, Staudt L, Martin R, McFarland HF. Expression profiling identifies responder and non-responder phenotypes to interferon-beta in multiple sclerosis. Brain 2003;126:1419-29

32. Weinstock-Guttman B, Badgett D, Patrick K, Hartrich L, Santos R, Hall D, Baier M, Feichter J, Ramanathan M. Genomic effects of IFN-beta in multiple sclerosis patients. J Immunol 2003;171:2694-702

33. Serrano-Fernández P, Möller S, Goertsches R, Fiedler H, Koczan D, Thiesen HJ, Zettl UK. Time course transcriptomics of IFNB1b drug therapy in multiple sclerosis. Autoimmunity. 2009 [Epub ahead of print]

34. Goertsches RH, Hecker M, Koczan D, Serrano-Fernandez P, Moeller S, Thiesen HJ, Zettl UK. Long-term genome-wide blood RNA expression profiles yield novel molecular response candidates for IFN-beta-1b treatment in relapsing remitting MS. Pharmacogenomics 2010;11:147-61.

35. Achiron A, Feldman A, Gurevich M.Molecular profiling of glatiramer acetate early treatment effects in multiple sclerosis. Dis Markers 2009;27:63-73

36. Wandinger KP, Sturzebecher CS, Bielekova B, Detore G, Rosenwald A, Staudt LM, McFarland HF, Martin R. Complex immunomodulatory effects of interferon-beta in multiple sclerosis include the upregulation of T helper 1-associated marker genes. Ann Neurol 2001;50:349–357

Chronic Viral Hepatitis and the Human Genome

Chloe L. Thio,[1] David L. Thomas,[1] and Mary Carrington[2]

Recently, molecular genetic tools have increased our ability to discover variations (polymorphisms) in the human genome and to correlate them with disease. Although still a discipline in its infancy, infectious disease outcomes may also be predetermined by genetic polymorphisms because differences in a host's susceptibility to infection and severity of disease cannot solely be attributed to the virulence of an organism. The earliest and best-known example of such genetically determined factors is malaria. Africans with Hemoglobin S are protected from severe infection with *Plasmodium falciparum* because the parasite cannot survive in the sickled cell.[1] Subsequently, tuberculosis and human immunodeficiency virus (HIV) outcomes have been correlated with polymorphisms in various genes.

Genetic associations may provide clues to fundamental questions about the pathogenesis of an organism and lead to new therapeutic avenues. For chronic viral hepatitis, this approach may help determine the basis for viral persistence and development of end-stage complications like cirrhosis or hepatocellular carcinoma. Initial genetic studies of viral hepatitis focused on human leukocyte antigen (HLA) associations. Much of this early literature is conflicting, possibly because of differences in study design and subjects, as well as failure to account for nonrandom association of genes (linkage disequilibrium). Currently, by exploiting the recent advances in molecular technology, such as genome-wide scans and large-scale polymorphism screening, investigators can screen a large number of subjects and identify multiple associated alleles. In the future, such studies may provide clues to help predict the natural outcome of a viral hepatitis infection or the response to therapy. This review provides some background into the molecular techniques used for genetic studies and summarizes the existing literature of host genetics with hepatitis B (HBV) and hepatitis C viruses (HCV).

CANDIDATE GENES

A candidate gene encodes for products that are hypothesized to influence the susceptibility to, or the course of, a particular disease. To determine if these genes contribute to the disease phenotype, its polymorphisms are examined in disease association or family studies. A candidate gene may be proposed if biological information suggests that its products may influence the known pathogenesis. Such an approach is illustrated by an example from the HIV literature. The chemokine receptor CCR5 was shown to behave as a coreceptor for HIV-1 entry into the CD4 cell,[2] so its gene was screened for polymorphisms. Individuals homozygous for a 32 base-pair (bp) deletion (CCR5Δ32) in this gene are resistant to HIV-1 infection, and those heterozygous for this deletion have delayed progression to AIDS.[3]

Other methods for proposing candidate genes include studying homologues of genes identified in an animal model or examining a gene implicated in a similar disease. For example, the gene NRAMP1 was examined for tuberculosis susceptibility because it is the human homologue of the mouse gene that is partly responsible for susceptibility to intracellular pathogens.[4] The hemochromatosis gene, HFE, is being investigated with viral hepatitis outcomes because iron has been implicated in viral hepatitis pathogenesis.[5] Limitations to these methods include prior identification of a gene, its polymorphisms, and its functional significance.

Another approach to identify candidate genes without these limitations is with a genome-wide scan, which uses densely distributed markers, either microsatellites or single nucleotide polymorphisms, to identify areas of interest in the genome. Distortions in the frequencies of these markers are compared between a cohort of patients with a specified outcome and those without the outcome or in a group of families with the disease phenotype.[6,7] If a disease-associated polymorphism and a marker are linked, then the marker will be found significantly more frequently in those with the disease.

Microsatellites are tandem repeats of two to four base pairs occurring approximately every 30 kilobases in the eukaryotic genome. In a study of five diseases (insulin-dependent diabetes mellitus, narcolepsy, uveitis, C2 deficiency, and multiple sclerosis) with known HLA associations, the microsatellites in closest proximity to the disease gene were significantly associated.[8] Thus, in diseases without previously recognized genetic associations, distortions in frequencies of microsatellite markers can focus the search for the associated gene.

Single nucleotide polymorphisms are polymorphisms produced by a single base change. These may be superior to microsatellites because they do not mutate as frequently, are more abundant, and are easier to analyze. If a large area of the genome is thought to be of interest, single nucleotide polymorphisms can be identified by denaturing high perfor-

---

Abbreviations: HIV, human immunodeficiency virus; HLA, human leukocyte antigen; HBV, hepatitis B virus; HCV, hepatitis C virus; AIDS, acquired immune deficiency syndrome; HFE, hemochromatosis gene; TNF-α, tumor necrosis factor-α; CTL, cytolytic T cell; PCR-SSP, polymerase chain reaction sequence-specific primers; PCR-SSO, polymerase chain reaction sequence-specific oligonucleotide; PCR-SSCP, polymerase chain recation sequence-specific conformational polymorphism; HBsAg, hepatitis B surface antigen; HBeAg, hepatitis B e antigen; anti-HBs, hepatitis B surface antibody; anti-HBe, hepatitis B e antibody; MHC, major histocompatibility complex; MBP, mannose binding protein.

From the [1]Division of Infectious Diseases, The Johns Hopkins School of Medicine, Baltimore and the [2]Laboratory of Genomic Diversity, National Cancer Institute, Frederick, MD.

Received August 9, 1999; accepted November 2, 1999.

This work was supported by NIH grants F32 DA 05887, R01 DA 13324, and U19 AI 40035 and by Public Health Service grant DA-005887. This project has been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, under Contract No. NO1-CO-56000. The publisher or recipient acknowledges right of the U.S. Government to retain a nonexclusive, royalty-free license in and to any copyright covering the article. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsements by the U.S. Government.

Address reprint requests to: Chloe L. Thio, M.D., Johns Hopkins University, Division of Infectious Diseases, 720 Rutland Avenue, Ross 1156, Baltimore, MD 21205. E-mail: cthio@jhmi.edu; fax: 410-614-9775.

This is a US government work. There are no restrictions on its use.

0270-9139/00/3104-0002$0.00/0
doi:10.1053/he.2000.4316

mance liquid chromatography, which is based on the property that heteroduplex DNA fragments compared with homoduplexes have reduced retention time on the liquid chromatography column.[9]

Major Histocompatibility Complex. The human major histocompatibility complex is located on the short arm of chromosome 6 and encodes for several protein products involved in immune function including complement, TNF-α, and the HLA complex, whose polymorphisms are often proposed as candidate genes (Fig. 1). The HLA complex consists of class I (HLA-A, -B, and -C) and class II (HLA-DRB1, -DQA1, -DQB1, -DPA1, and -DPB1) alleles. As the primary inducer of an immune response, the HLA present foreign antigens to both the CD4+ helper T cells and the CD8+ cytolytic T cells (CTLs). Class I molecules generally present antigens that are generated endogenously, including epitopes from viruses and other intracellular pathogens. To eliminate virally-infected hepatocytes, CTLs must recognize the combination of viral epitope and a class I antigen co-expressed on the surface of the hepatocyte. Class II molecules on antigen presenting cells present extracellularly derived antigens, including viral peptides, to CD4+ T cells to stimulate cytokine release generating both humoral and cell-mediated immune responses.

The HLA genes are the most diverse in the human genome, presumably an evolutionary adaptation to immune pressure from various infectious agents. Despite their heterogeneity, the HLA genes have been extensively characterized, and large-scale typing is now feasible. Because the methodology for class II typing has advanced more rapidly than for class I, most studies to date have employed serological class I typing, which is less specific than the recently-developed molecular techniques. Serological typing is based on HLA-specific antibody binding to antigens exposed on the cell surface. Thus, polymorphisms in the unexposed peptide-binding groove, which determines the antigenic epitopes presented to T cells, often escape detection. Presently, the most accurate methods for class I typing involve some combination of PCR-SSP (polymerase chain reaction sequence-specific primers), PCR-SSO (polymerase chain reaction sequence-specific oligonucleotide), and direct sequencing. In PCR-SSP, primers are designed to amplify one or more specific alleles by PCR, and the products are separated on an agarose gel.[10] Thus, determining the presence or absence of a product uniquely identifies a large number of alleles. In PCR-SSO, genomic DNA is amplified by PCR with locus-specific primers, the products are blotted onto a membrane, and the blot is then hybridized with oligonucleotide probes specific for particular alleles.[11] Accurate methods for class II HLA typing typically involve some combination of PCR-SSCP (sequence-specific conformational polymorphism), PCR-SSO, PCR-SSP, and sequencing. SSCP is based on the property that single base changes alter the conformation of single-stranded DNA, which leads to slower migration on a nondenaturing polyacrylamide gel.[12]

METHODS FOR CANDIDATE GENE ANALYSIS

Linkage. Linkage disequilibrium refers to the nonrandom association of alleles such that a particular allelic combination occurs more often than is expected by chance. For example, HLA-A1 and HLA-B8 occur in 27.5% and 15.7% of Caucasians, respectively. Thus, they would be expected to occur together in 4.3% (27.5% × 15.7%) of Caucasians assuming random mating, but they actually are found together twofold more often (9.8%) indicating they are in linkage disequilibrium and form a haplotype.[13] In family studies, linkage analysis is useful to study genes in large pedigrees. In association based studies, a marker locus, such as a microsatellite, is linked to the disease locus and can be compared in a case-control analysis.

Interpreting results of genetic studies can be complicated by linkage disequilibrium such that the implicated gene may be linked to the true gene of interest. This molecular confounding is similar to concluding that alcohol use causes lung cancer because of its epidemiological linkage to tobacco use. Linkage disequilibrium can occur through genetic proximity or from selection pressure favoring a particular combination of alleles at two distant loci.

Family Studies. Before the technology to study candidate genes in disease association studies was available, family studies were used, and are still used, to provide the first clues to a genetic basis for an infectious disease. In family studies, individuals with a particular disease comprise the initial cohort, and their relatives are studied to determine if they also have an excess frequency of the disease in question. Through the use of segregation analysis, these studies can suggest a pattern of inheritance of the susceptibility genes. With the advent of molecular genetics technology, results from disease association studies can be combined with family studies to determine the inheritance pattern of a particular candidate gene.

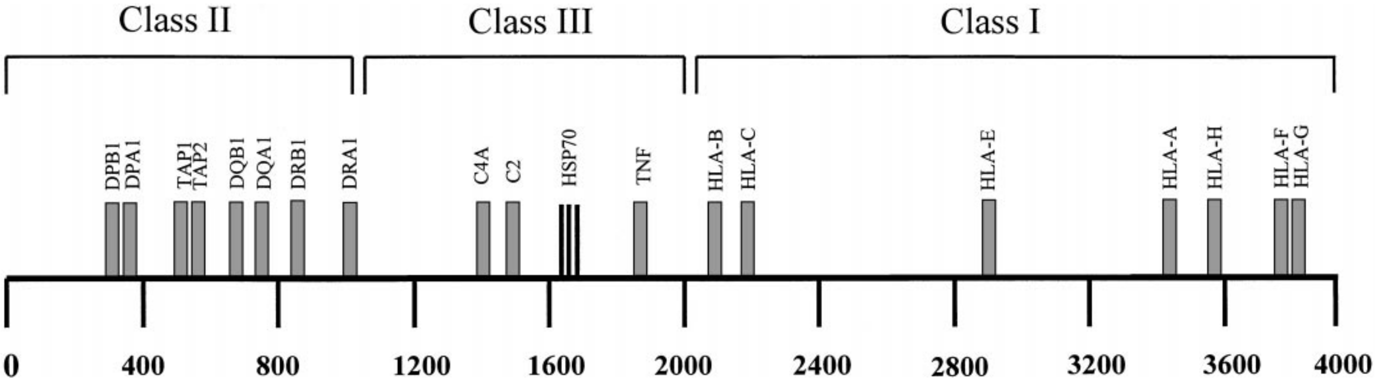


FIG. 1. The major histocompatibility complex (MHC) is located on the short arm of chromosome 6. Over 200 genes are encoded for in the MHC, many of which are involved in immune responsiveness. The relative locations of the major known genes are depicted in this schematic illustration.

15273350, 2000, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1053/he.2000.4316, Wiley Online Library on [15/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

In contrast to other family studies, twin and adoptee studies can separate the genetic and environmental components because monozygotic twins share 100% of genes and dizygotic twins share, on average, 50%. In general, family studies are relatively insensitive and require both uniform exposure to the infectious disease and high penetrance of the putative allele. Thus, in the majority of genetic studies of infectious diseases, a more sensitive way to detect disease-associated genes is to compare their allele frequencies between cases and controls.

Disease Association Studies. In disease association studies, a candidate gene polymorphism is studied in a group of individuals exposed to a particular infection. The frequencies of the polymorphism are then compared between those with and without a predefined study outcome. Using the example from HIV presented earlier, study outcomes with the CCR5 polymorphism included resistance versus susceptibility to infection and rapid versus slow progression to AIDS.[3] Most studies to date of HBV and HCV have employed this method of study, but difficulties involved in designing such studies include an inappropriate control group (detailed later) or a small number of study participants.[14-21]

The control group in most of the viral hepatitis studies is from the general population[22-28] or from blood or bone marrow donors,[29-31] which can lead to a misclassification bias because some of these individuals may not have been exposed to the virus. In such cases, true associations may be missed because unexposed subjects capable of becoming infected are included as controls. For studies of viral persistence, appropriate control groups are individuals that have cleared the infection and for studies of susceptibility to infection, the control group should be exposed and uninfected persons. The potential also exists for a control group to be enriched for another disease that has HLA or other genetic associations such as malaria, diabetes, HIV, etc.; thus, the associations discovered may reflect patterns with this other disease. Matching for such potential associations can eliminate this bias.

Genetic epidemiological studies require a large number of subjects to attain adequate statistical power. This problem is particularly apropos to HLA association studies because of the inordinately large number and fairly even distribution of alleles at some of these loci. As a hypothetical example, for an allele with a frequency of 10%, 307 case-control pairs are needed to have an 80% chance of detecting a twofold disease association. Most gene frequencies, especially in the HLA complex, are lower than 10%, further increasing the number of subjects required to investigate even modest disease associations.

Allele frequencies also vary between ethnic groups, so associations may only be detected in people of a particular ethnicity. The best example is the CCR5Δ32 polymorphism for HIV, which is found in Caucasians but is not detected in Africans.[32,33] Similarly, if a gene associated with viral hepatitis is limited to a particular ethnic group, then this association would not be detected in all studies; a multicohort study is ideal to examine such ethnically varied associations. For the same reason, the control population should be of the same ethnicity as the study population or alternatively, the two should be matched by race.

GENETICS AND HEPATITIS VIRUS

Hepatitis B Virus

Natural History. When hepatitis B virus (HBV) is acquired in adulthood, the majority of infections are cleared, with chronic infection occurring in 5% to 10% of cases.[34] However, 90% of childhood-acquired disease persists. In the developing world, where high endemic levels of HBV exist, acquisition usually occurs in childhood from either perinatal or horizontal (one child to another) transmission. In economically developed countries, HBV is usually acquired in adulthood via sexual transmission or injection drug use. When comparing studies of HBV and host genetics, the age-related variation in immune response may be a factor influencing the results.

Family Studies. Family studies in China provide some evidence that host genetic factors influence viral persistence, as higher concordance rates were found for HBeAg persistence in identical compared with nonidentical twins.[35] There are no family studies that have examined a genetic basis for susceptibility to HBV infection.

HLA Studies. The majority of host genetic studies with HBV have concentrated on HLA associations. Table 1 summarizes data from journal articles published between the years 1987 and 1998 reporting HLA class I and class II associations. All of these studies examined associations with persistence of HBV and a few studies additionally examined disease progression.[14,15,22-24,28-30,36-39] Seven of the twelve studies used controls who had cleared the infection,[14,15,28,29,36,37,39] and the remainder used controls from the general population. None of the seven identified associations between class I alleles and viral persistence, but only one used molecular typing.[14]

Class II associations were noted in four of the seven studies that compared viral persistence with viral clearance.[14,15,36,37] Using restriction fragment length polymorphism analysis, Thursz et al. examined a large cohort of children from The Gambia and found DRB1*1302 to be more frequent in the group that cleared the infection. The protective effect of this allele was validated in a smaller cohort of Gambian adults.[36] The cohort in both studies was a subset of a larger malaria study in which the DRB1*1302 allele was also found to be protective against severe malaria. The distribution of severe malaria cases in the persistent and self-limited HBV infections was not provided, so malaria may be a confounding factor in this study. In The Gambia, HBV is usually acquired in childhood via horizontal transmission. Because children are much more likely to become persistently infected than are adults, this association may not be generalizable to populations in which HBV is acquired in adulthood. The adults in the validation cohort probably also acquired their infection in childhood; thus a survivor bias may be present in the adult cohort. In support of this association, a separate German study designed to test the effect of DRB1*1302 in adults also suggested a protective role for this allele against viral persistence.[37]

A third study, which used SSCP and SSP, involved 31 self-limited and 60 persistent HBV infections in adult African-American injection-drug users from Baltimore, MD. This study showed the association of persistence with DQA1*0501 and DQB1*0301, which are found on a common DQA1-DQB1 haplotype. The 3 locus haplotype DQA1*0501-DQB1*0301-DRB1*1102 was also significantly associated with persistence.

15273350, 2000, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1053/he.2000.4316, Wiley Online Library on [15/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 1. Summary of All Statistically Significant HLA Associations With Hepatitis B Virus Persistence

Outcome (Numbers)	Controls (Numbers)	Risk Group	Population	Alleles Associated With Outcome	Comment	Ref
Persistence (31)	Clearance (60)	Adult, injection drug users	African-Americans	DQA*0501, DQB*0301; haplotype of these 2 alleles alone and with DRB1*1102	Small numbers	14
Persistence (185 children)	Clearance (218 children)	Children (≤10 yo)	Gambians	DRB1*1302 negatively associated	1) CI-serol typing 2) Distribution of malaria cases unknown 3) Childhood acquired disease	36
(40 adults)	195 (adults)	Adults	Gambians	DRB1*1302 negatively associated		
Persistence (21)	Clearance (31)	Adults; Blood donors, Hepatitis clinic	Qatarians	DR-7 positively associated; DR-2 negatively associated	1) Serol typing 2) Small numbers 3) Most controls unexposed 4) Clearance people not HBsAg tested	15
Persistence (107) (70 tested for DRB1)	Clearance (24) & Unexposed (101)	Adults; Liver center	Germans	DRB1*1301-2 negatively associated	1) CI serol typing 2) Most controls unexposed 3) No DQA, DQB testing 4) Clearance people not HBsAg tested	37
Persistence (97)	Clearance (52) & blood and marrow donors (489)	Adults	Italians		1) Serol typing 2) Most controls unexposed 3) Clearance people not HBsAg tested	29
Persistence (80)	Clearance (157)	Adults, HD patients	88% Caucasian 12% African		1) Serol typing 2) No DQA, DQB 3) Controls not HBsAg tested	28
Persistence and disease severity (22)	Blood donors (80)	Adults; Liver clinic	Caucasian		1) Controls unexposed 2) Small numbers 3) 17 DR3 and DR4 tested	30
Persistence (51)	Unspecified (97)	Adults	Latin Americans	B35 positively associated	1) Serol typing 2) DQA, DQB not tested 3) Controls unexposed 4) Small numbers	38
Disease severity and persistence (79-37 with CAH)	Clearance (117); random blood donors (200)	Adults	Dutch	DQw1 (↓ in CAH)	Serol typing	39
Disease severity, persistence (95 HBsAg healthy carriers, 210 chronic liver disease)	Random (253) (182-DR typing)	Adults	Taiwanese		1) Serol typing 2) Controls unexposed 3) DQA, DQB not tested	22
Persistence (78)	Random (340)	Adults	Japanese	Cw3 positively associated		23
Persistence (168)	Random (174)	Adults		A3, A11, B35, B51 positively associated	1) Controls unexposed 2) Serol typing 3) DQA, DQB not tested	24

Abbreviations: CI, class I; Serol, serological; CAH, chronic active hepatitis; HD, hemodialysis.

Lastly, a study of adults from Qatar detected different class II associations, but only 31 of the 131 controls were anti-HBs positive. Using serological class II typing, they showed DR7 to be a risk factor and DR2 to be protective. No difference in the frequency of DR6 (the supertype for the DRB1*1302 allele) or of DR4, which is in linkage disequilibrium with DQB1*0301, was observed.

Thus, one particular allele has not been clearly identified in association with HBV persistence. This ambiguity may be attributable to several possibilities including (1) a complex interaction involving multiple alleles, (2) differences among ethnic groups, (3) an association with a class I allele that has not been identified with molecular typing, or (4) an association with a gene in linkage disequilibrium with an HLA allele. Multicohort studies may be needed to clarify the data.

Theoretically, as first proposed by Zinkernagel and Doherty, heterozygosity at MHC loci could be advantageous for viral clearance since individuals who are fully heterozygous have twice as many alleles with which to present antigen than those who are fully homozygous.[40] For example, there are three classical class I loci (HLA-A, -B, -C), so an individual who is heterozygous at all loci has six different alleles with which to present antigen, but an individual who is fully homozygous has only three different alleles. Among HIV–infected individuals, those who are fully heterozygous for class I alleles are less likely to progress rapidly to AIDS.[41]

15273350, 2000, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1053/he.2000.4316, Wiley Online Library on [15/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Thursz and coworkers examined this in The Gambia and found significantly fewer individuals who were heterozygous for the class II alleles, DR and DQ, in the persistently infected compared with the transiently infected group.[42] This heterozygote advantage was also suggested in Thio and coworker's study but was not statistically significant.[14]

Three studies have examined the relationship between HLA and disease severity as determined by liver histology in persistently HBV–infected individuals.[22,30,39] None of these results have been reproduced.

The HLA association data thus far suggests that class II alleles may be more important in generating an effective immune response against HBV than class I alleles. The fact that antibodies to HBsAg are protective also implicates class II antigens as primary effectors against HBV infection. However, class I molecular typing has not been performed in a large study to date, so the role of class I awaits a more thorough investigation.

Other Candidate Genes. Tumor necrosis factor (TNF)-α, mannose binding protein (MBP), and the vitamin D receptor have been examined as potential candidate genes given that each of them may play a role in HBV pathogenesis. Patients chronically infected with HBV have increased levels of TNF-α and an upregulation of TNF-α receptors.[43,44] Also, in a HBV transgenic mouse model, CTL secretion of TNF-α is able to stop HBV gene expression.[45] Given these data, it is plausible that polymorphisms in the TNF-α gene may influence HBV persistence. This gene is located within the class III region of the MHC complex, and two polymorphisms in its promoter region, at positions -308 and -238, alter TNF-α expression.[46,47] The -238 promoter variant was significantly associated with HBV persistence in a cohort of 71 chronically and 32 transiently HBV–infected German patients.[48]

MBP affects inate immunity by activating the classical complement pathway and by acting as an opsonin binding directly to carbohydrate moieties on bacterial cell walls and to C1q receptors.[49,50] Several mutations in the MBP gene (codons 54, 57, and 52) and its promoter lead to low serum concentrations abolishing its ability to both activate complement and to act as an opsonin.[51] The HBV envlope contains a mannose-rich oligosaccharide to which MBP could potentially bind; thus these mutations have been investigated. One study compared a group with chronic HBV hepatitis (33 British Caucasians, 20 Asians, and 10 Others) to 98 British Caucasians and 123 Chinese controls with unknown exposure to HBV.[52] The codon 52 mutation of the MBP gene in British Caucasians, but not in the Asians, was associated with persistent HBV infection. Data from Chinese investigators supports this lack of assoiation between MBP polymorphisms and HBV persistence in Asians, but these investigators found the codon 54 mutation to be associated with symptomatic persistent infection.[53] In German Caucasians and in Gambians, these MBP polymorphisms were not asociated with chronic infection.[54,55]

In addition to its role in calcium regulation, the active form of vitamin D is an immunomodulatory hormone that inhibits the Th1 response and activates the Th2 response.[56] The vitamin D receptor gene has two known polymorphisms that have been found to be associated with tuberculosis, leprosy, and HIV.[57,58] Investigators in The Gambia have also studied these polymorphisms in relationship to HBV infection and found the tt genotype in one polymorphism to be associated with viral clearance.[57]

Discussion of HBV and Host Genetics. Thus far, the only reproducible association between host genetics and HBV infection involves HLA class II alleles, although the specific allele has varied between studies. If indeed specific alleles of the class II loci prove to be particularly efficient in viral clearance or slowing disease progression, then it follows that a CD4+ helper T cell response will be implicated as a critical factor in viral clearance. On the other hand, strong class I associations would suggest that a CD8+ CTL response is a critical factor in eliminating HBV. A primary role for class I involvement is a reasonable postulate because clearance of nonlytic viruses requires an efficient CTL response to eliminate the infected cells. A class I effect cannot be ruled out because few studies have addressed such a relationship and none have employed rigorous class I typing methodology using a large cohort. Further studies, including a multicohort study, are needed to clarify these preliminary associations and to investigate other potential candidate genes.

Hepatitis C Virus

Natural History. Hepatitis C virus is a bloodborne pathogen that is contracted in most cases via injection drug use and blood transfusions.[59] In contrast to adult-acquired HBV infection, the majority (75%-85%) of HCV infections persist. HCV antibody (anti-HCV) persists for years after HCV infection, even in those with self-limited infections. Thus, the presence of HCV RNA is used clinically to differentiate viral clearance from persistence. Because viremia can be intermittent in the first year of infection, demonstration of repeated presence or absence of HCV RNA is necessary to ascertain the outcome of an acute HCV infection.[60]

Genetic studies to date have examined 3 aspects of HCV infection: (1) clearance, (2) progression (cirrhosis), and (3) susceptibility to infection. The first, viral clearance, should be defined as a positive anti-HCV and two negative viral RNA tests at least six months apart; whereas, persistence of HCV implies continued HCV RNA positivity over a one-year period. In the second outcome, disease progression, the degree of liver damage can be semiquantitatively assessed by a system used to score liver biopsies, namely the histological activity index. Alternatively, clinical evidence of liver failure can be used. It is important to note that serious liver disease can be present without symptoms, potentially leading to misclassification of cases as controls. Thus, studies of HCV disease progression must include either large numbers of asymptomatic controls or those found to have limited disease on biopsy. Whether differential susceptibility to HCV infection actually occurs is not known because exposure is difficult to prove without an experimental model.

HLA Studies. Tables 2 through 4 outline statistically significant HLA associations from journal articles published between 1990 and 1998 for HCV outcomes of clearance, progression, and susceptibility to infection. These outcomes are discussed separately below.

Clearance. Three studies have examined class II associations with HCV clearance,[16-18] all of which defined clearance over a six-month period (Table 2). In these studies, DQB1*0301 was associated with clearance; Cramp et al. also identified DRB1*04 and DQA1*03 as protective alleles, which are in strong linkage disequilibrium with DQB1*0301. DRB1*1101, which is also in linkage disequilibrium with DQB1*0301, was associated with clearance in the Alric et al. study, and DRB1*11 was associated in the study by Minton

15273350, 2000, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1053/he.2000.4316, Wiley Online Library on [15/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 2. Summary of Statistically Significant HLA Associations With Hepatitis C Virus Clearance

Outcome (Numbers)	Controls (Numbers)	Risk Group	Population	Alleles Associated With Outcome	Comment	Ref
Clearance (49)	Persistent viremia (55)	IDU; Adults	Europeans	DQA1*03, DQB1*0301, DRB1*04 positively associated; DQB1*0302 negatively associated;	1) No class I 2) Small numbers 3) DQB1*0301 effect strongest; others linked	16
Clearance (25)	Persistent viremia (103)	Adult Transfusion 38%, IDU 24%	French, Caucasian	DQB1*0301 DRB1*1101 positively associated	1) No class I 2) No DQA 3) Small numbers 4) 100% linkage of DRB1*1101-DQB1*0301	17
Clearance (35)	Persistent viremia (135)	Blood donors	English	DQB1*0301, DRB1*11 positively associated	1) Effect strongest with DRB1*11 2) Small numbers	18

Abbreviation: IDU, injection drug use.

et al. The studies by Alric et al. and Cramp et al. suggest a role for DQB1*0301 in viral clearance and that two different DRB1 alleles were simply hitchhiking on DQB1*0301 in different populations. However, the study by Minton et al. suggests a stronger association for DRB1*11 than for DQB1*0301. None of these studies examined class I alleles.

Progression (Liver Failure). A number of additional studies have considered the association of HLA alleles with progression of liver disease[19-21,61,62] (Table 3). Two Japanese studies compared HCV carriers with normal liver function tests and normal histology with patients with elevated liver function tests and cirrhosis, respectively.[61,62] In both studies, DQB1*0401, DRB1*0405, and the two-locus haplotype consisting of these alleles were more frequent in those who developed chronic liver disease. However, because Kuzushita et al. also typed class I alleles, they showed that subjects with the three-locus haplotype, HLA-B54-DQB1*0401-DRB1*0405, were 13 times as likely to have liver injury relative to those with the two-locus haplotype without B54. In addition, a total of five class II alleles were suggested as protective against HCV disease progression between these two studies. Of the five, only DRB1*1302 showed the same trend for protection in these and two other Japanese studies.[19,21] DRB1*1101, which was associated with viral clearance,[17] was associated wih slower disease progression in two studies.[21,61] Additionally, Peano et al.[20] showed an inverse relationship between disease severity and frequency of DR5, which includes the allele DQB1*0301 associated with HCV clearance.[16,17]

Susceptibility to Infection. Several studies have attempted to examine susceptibility to HCV infection by comparing HLA alleles in individuals who are anti-HCV–positive with either the general population or with people who are anti-HCV–negative (Table 4).[25-29,31,63-65] Because the assumption cannot be made that the general population has been exposed to HCV, true associations may be missed by this experimental design. In addition, identification of individuals who have been exposed to HCV but remain seronegative is difficult.

One well-controlled study of the association of class II alleles with susceptibility to HCV infection compared 186 anti-HCV–positive patients, 121 of whom were HCV RNA–positive, to 328 anti-HCV–negative patients in a renal transplant clinic.[63] The primary mode of exposure was via multiple

TABLE 3. Summary of HLA Associations With Hepatitis C Virus Progression

Outcome (Numbers)	Controls (Numbers)	Risk Group	Population	Alleles Associated With Outcome	Comment	Ref
Normal ALT; Ab positive (33)	Elevated ALT (97)	Adults (⅓ transfusion)	Japanese	DRB1*1302, DRB1*1101, DQB1*0604 positively associated; DQB1*0401, DRB1*0405, B54 negatively associated	1) Serol class I 2) No DQA 3) Controls unexposed 4) B54-DQB1*0401-DRB1*0405 haplotype stronger association than without B54	61
No cirrhosis, chronic HCV (50)	Cirrhosis (67)	Adults	Japanese	DRB1*0901, DQB1*0303 positively associated; DRB1*0405, DQB1*0401 negatively associated	1) No class I, DQA 2) Small numbers 3) B54 frequency unknown	62
Normal ALT, chronic HCV (19)	Chronic active hepatitis (49)	Adults	Japanese	DR13	1) No class I, DQA, DQB 2) Small numbers	19
Asymptomatic, Ab positive (30)	Chronic liver disease (51)	Blood donors	Unknown	DR5	1) No class I typing 2) Ethnicity unknown 3) DQB1*0301 is in DR5 supertype 4) Small numbers	20
Asymptomatic chronic HCV (13)	Chronic liver disease (119)	Unknown	Japanese	DRB1*1101, 0802, 1302	1) No class I, DQA or DQB typing 2) Controls unexposed	21

Abbreviations: ALT, alanine transaminase; Ab, antibody; Serol, serological.

15273350, 2000, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1053/he.2000.4316, Wiley Online Library on [15/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

15273350, 2000, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1053/he.2000.4316, Wiley Online Library on [15/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 4. **Summary of HLA Associations With Susceptibility to HCV Infection**

Outcome (Numbers)	Controls (Numbers)	Risk Group	Population	Alleles Associated With Outcome	Comment	Ref
Chronic liver disease (108 with 76 Class II typed)	BM donors (181 with 105 Class II typed)	Adults	Parisians		1) Class I serol typing 2) Controls unexposed	31
HCV infected (186 with 121 RNA positive)	HCV uninfected (328)	Renal transplant clinic	Caucasian		1) Class I serol 2) Not clear if controls have same exposure	63
HCV infected (104 with 99 RNA positive)	Europeans typed previously (177)	HCV clinic, h/o IDU	European Caucasian	DQB1*0302 and DQA1*03 negatively associated	1) No class I 2) Controls unexposed	25
HCV infected (116)	HCV uninfected (30)	Thalassemics	Sardinians	DRB1*1601 DQB1*0502 negatively associated	1) No class I 2) Small number of uninfecteds	64
HCV infected (60)	Japanese (293)	Unknown	Japanese	B61, Cw3, DR4, DQB1*0401, 0402 positively associated DR9, DQB1*0301, 0303 negatively associated	1) Serol CI and DR typing 2) Controls random	26
HCV infected (64)	Blood donors (80)	Liver Center (transplant, IDU)	Caucasian		1) No class I, DQ 2) Controls unexposed	65
HCV infected (75)	Random (101)	Liver Center	Caucasian	DRB1*0301 positively associated DRB1*1301-2, DQA1*0103 (haplotype also) negatively associated	1) Class I-serological 2) Controls unexposed	27
HCV chronic liver disease (117)	BM donors (289)	Liver Center	Italians	B14 positively associated DR5 and DR5-DQ3 haplotype negatively associated	1) Class I and II serol typing 2) Controls unexposed	29
HCV infected (58)	Random (242)	Kidney transplant	88% French 12% African		1) Class I and II serol typing 2) Controls unexposed	28

Abbreviations: BM, bone marrow; Serol, serological; IDU, injection drug use.

blood transfusions, although the authors did not demonstrate an equivalent exposure rate in the two groups. Serological methods of HLA typing did not detect significant class I or class II differences between the groups. Congia et al. compared 116 anti-HCV–positive thalassemics to 30 anti-HCV–negative thalassemic controls and 606 ethnically similar controls.[64] The thalassemics had all been exposed to HCV through multiple blood transfusions. The haplotype DRB1*1601-DQB1*0502 was more frequent in the HCV–negative thalassemic controls relative to the anti-HCV–positive individuals. Several other studies have identified associations between class I and/or class II alleles and HCV infection,[25-29] but class I typing was performed using serological techniques, and controls always involved individuals who were unlikely to have been exposed to the virus. In general, among these studies, many different HLA associations were identified, but a few supported data from studies of HCV clearance and progression. Tibbs et al.[25] demonstrated that the allele DQA1*03, which was shown to favor clearance, was associated with protection against HCV infection. In a German study, DRB1*1302, which is associated with slower disease progression,[19,21] was shown to have decreased frequency in HCV–infected individuals compared with controls.[27] DR4 and DQB1*0401 were more frequent in Japanese HCV–infected subjects compared with random controls,[26] and in a Japanese study of progression, both of these alleles are associated with more rapid progression.[62]

A comparison of results in each of the three outcomes presents a wide array of HLA alleles that are associated with HCV infection. It is plausible that several HLA alleles may play a role in HCV infection or that these alleles are in linkage disequilibrium with another gene associated with HCV pathogenesis. Alternatively, these differences may be related to the small number of subjects and ethnic diversity of these studies and point to the need for a large, multi-ethnic confirmatory study.

Other Gene Markers. The two other candidate genes that have been studied in connection with HCV persistence are TNF-α and the hemochromatosis gene, HFE. Mutations in the TNF-α gene have been proposed as candidate genes since TNF-α levels are elevated in HCV infection perhaps because of production by CTLs in the liver.[66] In a study of 82 persistently HCV–infected patients and 99 controls, the -238A promoter variant was significantly associated with persistent HCV infection.[67]

HFE was recently discovered and is beginning to be studied in HCV infection because of the potential affect of iron on HCV pathogenesis. Iron may promote HCV replication, increase oxidative stress, deplete hepatoprotective factors, or impair host immunity. HFE is a MHC-class-I–like gene located on chromosome 6p telomeric to the HLA-A complex, which has two identified mutations. The first is at amino acid position 282 (C282Y) and the second is at amino acid position 63 (H63D). It is thought that the protein encoded by the HFE gene may directly limit the amount of iron absorbed in the small intestine or may regulate gene products that control iron absorption; therefore, these mutations may lead to increased iron absorption.[68]

A few studies have investigated the role of these mutations in individuals with chronic HCV infection. Hezode and coworkers found no difference in the frequency of heterozygosity for the C282Y mutation between HCV–infected patients with chronic active hepatitis and the general population.[69] Kazemi-Shirazi and coworkers compared both mutations between chronically HCV–infected patients and random controls.[70] There was no difference in the frequency of either mutation between the groups, however, iron overload was more common in patients with an HFE gene mutation. These results led them to conclude that the C282Y mutation contributes to but does not fully explain iron accumulation in chronically HCV–infected persons. Smith et al. studied 137 people with chronic HCV hepatitis and concluded that individuals heterozygous for the C282 mutation have more fibrosis but only minor increases in iron stores.[71] Overall, the data do not convincingly show a role between HFE mutations and HCV disease progression, although the studies to date have been small.

Discussion of Host Genetics and HCV. The literature on the host genome and HCV presents some reproducible HLA associations although further study is needed to convincingly identify alleles that affect response to infection and disease outcome. DQB1*0301 is associated with clearance[16,17] and with less severe disease,[20] but a more rigorous analysis of the other class II and class I alleles is essential to determine whether DQB1*0301 is simply hitchhiking with the actual protective allele at another HLA locus or another gene. As in HBV, class I associations have not been examined with rigorous class I typing methodology, and examination of candidate genes has just begun.

SUMMARY

The immunogenetics of chronic viral hepatitis is an immature field, but the molecular tools are becoming available that will allow a more thorough exploration of this area. Inconsistencies in the data presented may be related to inaccurate and/or low resolution HLA typing methods, linkage disequilibrium between genes, ethnic differences in the populations studied, and small sample sizes. In addition, comparison of HBV studies is complicated by differences in immune responses depending on age at acquisition of infection. Genetic associations found in some of these studies support the hypothesis that there is a genetic basis for viral hepatitis outcomes, but a multi-cohort study may be required to overcome the limitations of previous studies.

Finding an association between a gene and a viral hepatitis outcome does not imply causality, but instead it opens new avenues to explore and to understand the pathogenesis of these viruses. Thus, such genetic studies can only begin to answer questions about the factors that influence the outcome of an acute viral hepatitis infection. Further studies based on this information include finding a medical basis for the association and embarking on functional studies of the implicated genes. Such studies can provide clues to the pathogenesis of both HBV and HCV and may provide a rationale on which to base future therapeutic strategies.

REFERENCES

1. Strickland GT: Malaria. In: Strickland GT, ed. Hunter's Tropical Medicine. Philadelphia: W.B. Saunders Company, 1991;586-602.
2. Deng H, Liu R, Ellmeier W, Choe S, Unutmaz D, Burkhart M, Di Marzio P, et al. Identification of a major co-receptor for primary isolates of HIV-1. Nature 1996;381(6584):661-666.
3. Dean M, Carrington M, Winkler C, Huttley GA, Smith MW, Allikmets RGJ, Goedert JJ, et al. Genetic restriction of HIV-1 infection and progression to AIDS by a deletion allele of the CKR5 structural gene. Hemophilia Growth and Development Study, Multicenter AIDS Cohort Study, Multicenter Hemophilia Cohort Study, San Francisco City Cohort, ALIVE Study. Science 1996;273(5283):1856-1862.
4. Bellamy R, Ruwende C, Corrah T, McAdam KP, Whittle HC, Hill AV. Variations in the NRAMP1 gene and susceptibility to tuberculosis in West Africans. N Engl J Med 1998;338(10):640-644.
5. Bonkovsky HL, Banner BF, Rothman AL. Iron and chronic viral hepatitis. HEPATOLOGY 1997;25(3):759-768.
6. Dib C, Faure S, Fizames C, Samson D, Drouot N, Vignal A, et al. A comprehensive genetic map of the human genome based on 5,264 microsatellites. Nature 1996;380(6570):152-154.
7. Liu W, Smith DI, Rechtzigel KJ, Thibodeau SN, James CD. Denaturing high performance liquid chromatography (DHPLC) used in the detection of germline and somatic mutations. Nucleic Acids Res 1998;26(6): 1396-1400.
8. Carrington M, Marti D, Wade J, Klitz W, Barcellos L, Thomson G, Chen J, et al. Microsatellite markers in complex disease: mapping disease-associated regions within the human major histocompatibility complex. In: Goldstein DB and Schlotterer C, eds. Microsatellites: Evolution and Applications. Oxford: Oxford University Press, 1999;225-237.
9. Giordano M, Oefner PJ, Underhill PA, Cavalli SL, Tosi R, Richiardi PM. Identification by denaturing high-performance liquid chromatography of numerous polymorphisms in a candidate region for multiple sclerosis susceptibility. Genomics 1999;56(3):247-253.
10. Tonks S, Marsh SGE, Bunce M, Moses JH, Karusa P, Sadler AM, Petronzelli F, et al. HLA Class I DNA typing study. In: Charron D, ed. Genetic diversity of HLA: functional and medical implications. Paris: EDK Medical and Scientific International Publisher, 1997;199-215.
11. Gao X, Jakobsen IB, Serjeantson SW. Characterization of the HLA-A polymorphism by locus-specific polymerase chain reaction amplification and oligonucleotide hybridization. Hum Immunol 1994;41:267-279.
12. Carrington M, Millter T, White Me, Gerrard B, Stewart C, Dean M, Mann D. Typing of HLA-DQA1 and DQB1 using DNA single-stranded conformational polymorphism. Hum Immunol 1992;33:208-212.
13. Trowsdale J. Molecular genetics of HLA class I and class II regions. In: Browning M, McMichael A, eds. HLA and MHC: genes, molecules, and function. Oxford: BIOS Scientific Publishers, Ltd, 1996;23-37.
14. Thio CL, Carrington M, O'Brien SJ, Vlahov D, Nelson KE, Astemborski JA, Thomas DL. The association of HLA alleles and clearance of hepatitis B among African Americans. J Infect Dis 1999;179(4):1004-1006.
15. Almarri A, Batchelor JR. HLA and hepatitis B infection. Lancet 1994;344: 1194-1195.
16. Cramp ME, Carucci P, Underhill J, Naoumov NV, Williams R, Donaldson PT. Association between HLA class II genotype and spontaneous clearance of hepatitis C viraemia. J Hepatol 1998;29(2):207-213.
17. Alric L, Fort M, Izopet J, Vinel JP, Charlet JP, Selves J, Puel J, et al. Genes of the major histocompatibility complex class II influence the outcome of hepatitis C virus infection. Gastroenterology 1997;113(5):1675-1681.
18. Minton EJ, Smillie D, Neal KR, Irving WL, Underwood JCE, James V, and members of the Trent Hepatitis C Virus Study Group. Association between MHC class II alleles and clearance of circulating hepatitis C virus. J Infect Dis 1998;178(1):39-44.
19. Kuzushita N, Hayashi N, Katayama K, Hiramatsu N, Yasumaru M, Murata H, Shimizu Y, et al. Increased frequency of HLA DR13 in hepatitis C virus carriers with persistently normal ALT levels. J Med Virol 1996;48(1):1-7.
20. Peano G, Menardi G, Ponzetto A, Fenoglio LM. HLA-DR5 antigen: a genetic factor influencing the outcome of hepatitis C virus infection. Arch Int Med 1994;154:2733-2736.
21. Yasunami R, Miyamoto T, Kanda T. HLA-DRB1 is related to pathological changes of the liver in chronic hepatitis C. Hepatol Res 1997;7:3-12.
22. Yang P-M, Sung JL, Chen DS. HLA-A,B,C and DR antigens in chronic hepatitis B viral infection. Hepato-Gastroenterology 1989;36:363-366.
23. Watanabe F, Zeniya M, Aizawa Y, Andoh H, Ohkoshi H, Miyazaki H, Kameda H, et al. Genetic studies on hepatitis B virus carriers. Acta Hepatol Jpn 1989;30:614-621.
24. Haciberktasoglu A, Barut A, Inal A. HLA-A,B,C, and DR antigens in chronic hepatitis B infection. Mikrob Bull 1993;27:241-248.
25. Tibbs C, Donaldson P, Underhill JA, Thomson L, Manabe K, Williams R. Evidence that the HLA DQA1*03 allele confers protection from chronic HCV-infection in northern european caucasoids. HEPATOLOGY 1996;24: 1342-1345.
26. Higashi Y, Kamikawaji N, Suko H, Ando M. Analysis of HLA alleles in

15273350, 2000, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1053/he.2000.4316, Wiley Online Library on [15/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Japanese patients with cirrhosis due to chronic hepatitis C. Journal of Gastro & Hepatol 1996;11(3):241-246.

27. Hohler T, Gerken G, Notghi A, Knolle P, Lubjuhn R, Taheri H, Schneider PM, et al. MHC class II genes influence the susceptibility to chronic active hepatitis C. J Hepatol 1997;27:259-264.
28. Verdon R, Pol S, Landais P, Mattlinger B, Carnot F, Brechot C, Busson M, et al. Absence of association between HLA antigens and chronicity of viral hepatitis in haemodialyzed patients. J Hepatol 1994;21(3):388-393.
29. Zavaglia C, Bortolon C, Ferrioli G, Rho A, Mondazzi L, Bottelli RGA, Ghessi A, et al. HLA typing in chronic type B, D and C hepatitis. J Hepatol 1996;24(6):658-665.
30. Czaja AJ, Carpenter HA, Santrach PJ, Moore SB. Significance of human leukocyte antigens DR3 and DR4 in chronic viral hepatitis. Dig Dis Sci 1995;40(10):2098-2106.
31. Vitte RL, Fortier C, Richardet JP, Grimbert S, Trinchet JC, Beaugrand M, Le Page V, et al. HLA antigens in patients with chronic hepatitis C. Tissue Antigens 1995;45(5):356-361.
32. Libert F, Cochaux P, Beckman G, Samson M , Aksenova M, Cao A, et al. The deltaccr5 mutation conferring protection against HIV-1 in Caucasian populations has a single and recent origin in Northeastern Europe. Hum Mol Genet 1998;7(3):399-406.
33. Martinson JJ, Chapman NH, Rees DC, Liu YT, Clegg JB. Global distribution of the CCR5 gene 32-basepair deletion. Nat Genet 1997; 16(1):100-103.
34. Lee WM. Hepatitis B virus infection. N Engl J Med 1997;337(24):1733-1745.
35. Lin TM, Chen CJ, Wu MM. Hepatitis B virus markers in chinese twins. Anticancer Research 1989;9:737-742.
36. Thursz MR, Kwiatkowski D, Allsopp CEM, Greenwood BM, Thomas HC, Hill AVS. Association between an MHC class II allele and clearance of hepatitis B virus in the Gambia. N Engl J Med 1995;332:1065-1069.
37. Hohler T, Gerken G, Notghi A, Lubjuhn R, Taheri H, Protzer U, Lohr HF, et al. HLA-DRB1*1301 and *1302 protect against chronic hepatitis B. J Hepatol 1997;26:503-507.
38. Mota AH, Fainboim H, Terg R, Fainboim L. Association of chronic active hepatitis and HLA B35 in patients with hepatitis B virus. Tissue Antigens 1987;30:238-240.
39. Van Hattum J, Schreuder GM, Schalm SW. HLA antigens in patients with various courses after hepatitis B virus infection. HEPATOLOGY 1987;7(1): 11-14.
40. Doherty PC, Zinkernagel RM. Enhanced immunological surveillance in mice heterozygous at the H-2 gene complex. Nature 1975;256(5512): 50-52.
41. Carrington M, Nelson GW, Martin MP, Kissner T, Vlahov D, Goedert JJ, Kaslow R, et al. HLA and HIV-1: heterozygote advantage and B*35-Cw*04 disadvantage. Science 1999;283(5408):1748-1752.
42. Thursz M, Thomas HC, Greenwood BM, Hill AVS. Heterozygote advantage for HLA class-II type in hepatitis B virus infection. Nature Genetics 1997;17:11-12.
43. Sheron N, Lau J, Daniels H, Goka J, Eddleston A, Alexander GJ, Williams R. Increased production of tumour necrosis factor alpha in chronic hepatitis B virus infection. J Hepatol 1991;12(2):241-245.
44. Marinos G, Naoumov NV, Rossol S, Torre F , Wong PY, Gallati H, Portmann B, et al. Tumor necrosis factor receptors in patients with chronic hepatitis B virus infection. Gastroenterology 1995;108(5):1453-1463.
45. Guidotti LG, Ishikawa T, Hobbs MV, Matzke B, Schreiber R, Chisari FV. Intracellular inactivation of the hepatitis B virus by cytotoxic T lymphocytes. Immunity 1996;4(1):25-36.
46. Wilson AG, de Vries N, Pociot F, di Giovine FS, van der Putte LB, Duff GW. An allelic polymorphism within the human tumor necrosis factor alpha promoter region is strongly associated with HLA A1, B8, and DR3 alleles. J Exp Med 1993;177(2):557-560.
47. D'Alfonso S, Richiardi PM. A polymorphic variation in a putative regulation box of the TNFA promoter region. Immunogenetics 1994; 39(2):150-154.
48. Hohler T, Kruger A, Gerken G, Schneider PM, Meyer zBK, Rittner C. A tumor necrosis factor-alpha (TNF-alpha) promoter polymorphism is associated with chronic hepatitis B infection. Clin Exp Immunol 1998;111(3):579-582.
49. Kuhlman M, Joiner K, Ezekowitz RA. The human mannose-binding protein functions as an opsonin. J Exp Med 1989;169(5):1733-1745.
50. Matsushita M, Fujita T. Activation of the classical complement pathway by mannose-binding protein in association with a novel C1s-like serine protease. J Exp Med 1992;176(6):1497-1502.
51. Summerfield JA, Ryder S, Sumiya M, Thursz M, Gorchein A, Monteil MA, Turner MW. Mannose binding protein gene mutations associated with unusual and severe infections in adults. Lancet 1995;345(8954):886-889.
52. Thomas HC, Foster GR, Sumiya M, McIntosh D, Jack DL, Turner MW, Summerfield JA. Mutation of gene of mannose-binding protein associated with chronic hepatitis B viral infection. Lancet 1996;348(9039):1417-1419.
53. Yuen MF, Lau CS, Lau YL, Wong WM, Cheng CC, Lai CL. Mannose binding lectin gene mutations are associated with progression of liver disease in chronic hepatitis B infection. HEPATOLOGY 1999;29(4):1248-1251.
54. Hohler T, Wunschel M, Gerken G, Schneider PM, Meyer zBK, Rittner C. No association between mannose-binding lectin alleles and susceptibility to chronic hepatitis B virus infection in German patients. Exp Clin Immunogenet 1998;15(3):130-133.
55. Bellamy R, Ruwende C, McAdam KP, Thursz M, Sumiya M, Summerfield J, Gilbert SC, et al. Mannose binding protein deficiency is not associated with malaria, hepatitis B carriage nor tuberculosis in Africans. Q J Med 1998;91(1):13-18.
56. Long KZ, Santos JI. Vitamins and the regulation of the immune response. Pediatr Infect Dis J 1999;18(3):283-290.
57. Bellamy R, Ruwende C, Corrah T, McAdam KP, Thursz M, Whittle HC, Hill AV. Tuberculosis and chronic hepatitis B virus infection in Africans and variation in the vitamin D receptor gene. J Infect Dis 1999;179(3): 721-724.
58. Hill AV. The immunogenetics of human infectious diseases. Annu Rev Immunol 1998;16:593-617.
59. Anonymous. Recommendations for prevention and control of hepatitis C virus (HCV) infection and HCV-related chronic disease. MMWR 1998;47(RR-19):1-39.
60. Villano SA, Vlahov D, Nelson KE, Cohn S, Thomas DL. Persistence of viremia and the importance of long-term follow-up after acute hepatitis C infection. HEPATOLOGY 1999;29(3):908-914.
61. Kuzushita N, Hayashi N, Moribe T, Katayama K, Kanto T, Nakatani S, Kaneshige T, et al. Influence of HLA haplotypes on the clinical courses of individuals infected with hepatitis C virus. HEPATOLOGY 1998;27(1):240-244.
62. Aikawa T, Kojima M, Onishi H, Tamura R, Fukuda S, Suzuki T, Tsuda F, et al. HLA DRB1 and DQB1 alleles and haplotypes influencing the progression of hepatitis C. J Med Virol 1996; 49(4):274-278.
63. Chen DF, Endres W, Kliem V, Tillmann HL, Brunkhorst R, Koch KM, Manns MP, et al. No significant influence of HLA determinants on susceptibility to hepatitis C virus infection in Caucasian patients with end-stage renal disease. Liver 1996;16(6):384-389.
64. Congia M, Clemente MG, Dessi C, Cucca F, Mazzoleni AP, Frau F, Lampis R, et al. HLA class II genes in chronic hepatitis C virus-infection and associated immunological disorders. HEPATOLOGY 1996;24(6):1338-1341.
65. Czaja AJ, Carpenter H, Santrach PJ, Moore SB. DR human leukocyte antigens and disease severity in chronic hepatitis C. J Hepatol 1996;24(6): 666-673.
66. Koziel MJ, Dudley D, Afdhal N, Grakoui A, Rice CM, Choo QL, Houghton M, et al. HLA class I-restricted cytotoxic T lymphocytes specific for hepatitis C virus - Identification of multiple epitopes and characterization of patterns of cytokine release. J Clin Invest 1995;96(5): 2311-2321.
67. Hohler T, Kruger A, Gerken G, Schneider PM, Meyer zBK, Rittner C. Tumor necrosis factor alpha promoter polymorphism at position -238 is associated with chronic active hepatitis C infection. J Med Virol 1998;54(3):173-177.
68. Feder JN, Gnirke A, Thomas W, Tsuchihashi Z, Ruddy DA, Basava A, Dormishian F, et al. A novel MHC class I-like gene is mutated in patients with hereditary haemochromatosis. Nature Genetics 1996;13(4):399-408.
69. Hezode C, Cazeneuve C, Coue O, Pawlotsky JM, Zafrani ES, Amselem S, Dhumeaux D. Hemochromatosis Cys282Tyr mutation and liver iron overload in patients with chronic active hepatitis C. HEPATOLOGY 1998;27(1):306.
70. Kazemi-Shirazi L, Datz C, Maier-Dobersberger T, Kaserer K, Hackl F, Polli C, Steindl PE, et al. The relation of iron status and hemochromatosis gene mutations in patients with chronic hepatitis C. Gastroenterology 1999;116:127-134.
71. Smith BC, Grove J, Guzail MA, Day CP, Daly AK, Burt AD, Bassendine MF. Heterozygosity for hereditary hemochromatosis is associated with more fibrosis in chronic hepatitis C. HEPATOLOGY 1998;27(6):1695-1699.

15273350, 2000, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1053/he.2000.4316, Wiley Online Library on [15/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Host genetic effects in pneumonia

Hung-Hsin Chen,[1] Douglas M. Shaw,[1] Lauren E. Petty,[1] Misa Graff,[2] Ryan J. Bohlender,[3] Hannah G. Polikowsky,[1] Xue Zhong,[1] Daeeun Kim,[2] Victoria L. Buchanan,[2] Michael H. Preuss,[4] Megan M. Shuey,[1] Ruth J.F. Loos,[4] Chad D. Huff,[3] Nancy J. Cox,[1] Julie A. Bastarache,[5] Lisa Bastarache,[6] Kari E. North,[2] and Jennifer E. Below[1,*]

Summary

Given the coronavirus disease 2019 (COVID-19) pandemic, investigations into host susceptibility to infectious diseases and downstream sequelae have never been more relevant. Pneumonia is a lung disease that can cause respiratory failure and hypoxia and is a common complication of infectious diseases, including COVID-19. Few genome-wide association studies (GWASs) of host susceptibility and severity of pneumonia have been conducted. We performed GWASs of pneumonia susceptibility and severity in the Vanderbilt University biobank (BioVU) with linked electronic health records (EHRs), including Illumina Expanded Multi-Ethnic Global Array (MEGA^{EX})-genotyped European ancestry (EA, n= 69,819) and African ancestry (AA, n = 15,603) individuals. Two regions of large effect were identified: the *CFTR* locus in EA (rs113827944; OR = 1.84, p value = 1.2×10^{-36}) and *HBB* in AA (rs334 [p.Glu7Val]; OR = 1.63, p value = 3.5×10^{-13}). Mutations in these genes cause cystic fibrosis (CF) and sickle cell disease (SCD), respectively. After removing individuals diagnosed with CF and SCD, we assessed heterozygosity effects at our lead variants. Further GWASs after removing individuals with CF uncovered an additional association in *R3HCC1L* (rs10786398; OR = 1.22, p value = 3.5×10^{-8}), which was replicated in two independent datasets: UK Biobank (n = 459,741) and 7,985 non-overlapping BioVU subjects, who are genotyped on arrays other than MEGA^{EX}. This variant was also validated in GWASs of COVID-19 hospitalization and lung function. Our results highlight the importance of the host genome in infectious disease susceptibility and severity and offer crucial insight into genetic effects that could potentially influence severity of COVID-19 sequelae.

The COVID-19 pandemic is a serious threat to public health; over 58 million confirmed cases from 191 countries have been reported (see "Coronavirus Disease 2019" and "COVID-19 Map" in Web resources). Pneumonia is a common complication of COVID-19 and may lead to acute respiratory distress syndrome (ARDS) and death.[1] In the context of the COVID-19 pandemic, identifying factors that influence host susceptibility to and severity of pneumonia has never been more important. Several clinical factors and underlying conditions that influence susceptibility to and severity of community-acquired pneumonia in adults have been identified, including age, chronic bronchitis or chronic obstructive pulmonary disease (COPD), asthma, obesity, diabetes, and others.[2] These risk factors have also been observed in COVID-19-associated pneumonia.[3–6] However, little is known about the role of the host genome in the susceptibility to and severity of pneumonia.

There is a paucity of studies interrogating host genetic susceptibility to and severity of pneumonia. While previous studies identified suggestive associations with childhood pneumonia, survival from sepsis due to pneumonia, and severe pneumonia following influenza A/H1N1 infection,[7–9] three genome-wide significant associations with pneumonia, one in the *HLA* class I region, have been mapped,[10] and two additional genome-wide significant independent hits on chromosome 15 were identified in a meta-analysis of the UK Biobank and FinnGen.[11] To identify additional genetic loci impacting pneumonia in existing data resources, we aimed to identify genetic variants associated with susceptibility to and severity of pneumonia by leveraging electronic health records (EHRs) from a large-scale biobank.

The Vanderbilt University Medical Center biobank (BioVU) comprises over 110,000 participants with linked EHRs genotyped on the Illumina Expanded Multi-Ethnic Genotyping Array (MEGA^{EX}) or another genome-wide array. We defined pneumonia cases by using 81 pneumonia-related diagnosis codes (Supplemental box) from the International Classification of Diseases Ninth Revision, Clinical Modification (ICD-9 CM) and used the subjects without any diagnosis of pneumonia as population-based controls. We used hospitalization status as a proxy for severity, determined via relevant Current Procedural Terminology codes recorded within 5 days of pneumonia diagnosis (Supplemental box). In 69,819 MEGA^{EX}-genotyped European ancestry (EA) individuals, 8,889 individuals with pneumonia were identified, including 5,774 with pneumonia-associated hospitalization (inpatients). In 15,603 MEGA^{EX}-genotyped African ancestry (AA) individuals, we

---

[1]Vanderbilt Genetics Institute and Division of Genetic Medicine, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN 37232, USA; [2]Department of Epidemiology, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27516, USA; [3]Department of Epidemiology, The University of Texas MD Anderson Cancer Center, Houston, TX 77230, USA; [4]The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA; [5]Allergy, Pulmonary and Critical Care Medicine, Vanderbilt University School of Medicine, Nashville, TN 37232, USA; [6]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN 37232, USA
*Correspondence: jennifer.e.below@vanderbilt.edu

https://doi.org/10.1016/j.ajhg.2020.12.010.

© 2020

Table 1. Significant findings from discovery pneumonia GWAS, replication, and COVID-19 validation

					BioVU MEGA^{EX}		BioVU non-MEGA^{EX}		UK Biobank[a]		COVID-19 HGI[b]	
Gene	**SNP**	**Chr**	**Pos**	**Pop**	**OR**	**p value**	**OR**	**p value**	**OR**	**p value**	**OR**	**p value**
Susceptibility to pneumonia/COVID-19												
CFTR	rs113827944	7	117,299,434	EA1	1.84	1.20E−36	1.92	2.42E−06	1.04	4.90E−01	1.14	1.46E-01
HBB	rs334	11	5,248,232	AA1	1.63	3.50E−13	1.15	5.37E−01	0.89	7.37E−01	N/A	N/A
UQCRFS1	rs148218440	19	29,589,778	AA2	2.21	3.60E−08	0.70	4.27E−01	2.23	1.43E−01	N/A	N/A
Severity of pneumonia/COVID-19												
CFTR	rs113827944	7	117,299,434	EA2	1.70	4.01E−09	1.92	2.70E−02	1.04	4.90E−01	1.66	7.59E-02
R3HCC1L	rs10786398	10	99,926,570	EA3	1.22	3.50E−08	1.01	9.50E−01	1.02	2.48E−01	1.19	3.72E-02
R3HCC1L	rs884811	10	99,923,763	EA3	1.16	4.80E−06	1.10	3.64E−01	1.05	3.49E−03	N/A	N/A
R3HCC1L	rs7086391	10	100,013,563	AA3	1.42	1.10E−03	0.85	5.96E−01	1.10	6.73E−01	N/A	N/A
HBB	rs334	11	5,248,232	AA4	2.57	4.50E−12	1.73	2.96E−01	0.89	7.37E−01	N/A	N/A

EA1, European individuals with pneumonia versus control individuals; EA2, European pneumonia inpatients versus outpatients; EA3, European pneumonia inpatients versus outpatients, individuals with CF removed; AA1, African American individuals with pneumonia versus control individuals; AA2, African American pneumonia inpatients versus matched control individuals, individuals with CF and SCD removed; AA3, African American pneumonia inpatients versus outpatients, individuals with CF and SCD removed; AA4, African American pneumonia inpatients versus outpatients; Chr, chromosome; Pos, position; Pop, population.

[a] In UK Biobank, only the comparison between individuals with pneumonia versus the general population is available.

[b] In COVID-19 HGI, we used "C2" phenotypes, which compared individuals with COVID-19 to population controls, as susceptibility and "B1" results, which compared hospitalized individuals with COVID-19 to non-hospitalized control individuals with COVID-19, as severity.

identified 1,710 individuals with pneumonia, of which 1,043 were inpatients. In both EA and AA subjects, we observed a significantly higher prevalence of obesity, COPD, diabetes, asthma, and liver and renal diseases in subjects with a pneumonia diagnosis (Tables S1 and S2). We sought replication of our top findings in two independent datasets: UK Biobank ($n = 451{,}305$) and 7,985 non-overlapping BioVU subjects, who were genotyped on arrays other than MEGA^{EX} (non-MEGA^{EX}) (Tables S3–S5).

Genetic imputation in MEGA^{EX}-genotyped subjects was conducted with minimac4 on the Michigan Imputation Server[12] with a reference panel of Haplotype Reference Consortium r1.1. A total of 39,635,008 SNPs was imputed. In EA, only 7,167,360 SNPs with an imputation info score greater than 0.4 and minor allele frequency (MAF) greater than 1% were used for further GWAS and GReX imputation. In AA, 13,633,982 SNPs passed quality control filter. A more stringent MAF cutoff was applied in the comparison of AA pneumonia inpatient versus outpatient because of the smaller sample size ($n = 1{,}710$; 7,594,451 variants with MAF $> 5\%$ and imputation info score > 0.4). Because of the existence of genetic relatedness in BioVU, we utilized a generalized estimating equations (GEE) model to perform GWASs with SUGEN.[13] Since SUGEN requires known family networks representing close relatedness within a dataset, we used PRIMUS to reconstruct non-directional family networks, including all first- and second-degree,[14] and we used ERSA to verify the families with more than five members.[15] Among MEGA^{EX}-genotyped subjects, 5,019 families (size 2 or greater) were identified in BioVU EAs and 1,699 families in BioVU AAs. We included age, sex, and three principal components to capture genetic ancestry as covariates in the association tests.

Separate GWASs for susceptibility and severity of pneumonia were performed in EA and AA, each of which identified a major locus. In EA, rs113827944 (MAF $=$ 2.1%) was significantly associated with both pneumonia susceptibility and severity (affected individuals versus control individuals, odds ratio [OR] $= 1.84$, p value $= 1.2 \times 10^{-36}$; inpatients versus outpatients, OR $= 1.70$, p value $= 4.0 \times 10^{-9}$). Replication of this lead SNP was observed in BioVU non-MEGA^{EX}-genotyped EA (affected individuals versus control individuals, OR $= 1.92$, p value $= 2.4 \times 10^{-6}$). Another SNP, rs334 (p.Glu7Val, MAF $=$ 5.8%) was significantly associated with both susceptibility and severity in AA (affected individuals versus control individuals, OR $= 1.63$, p value $= 3.5 \times 10^{-13}$; inpatients versus outpatients, OR $= 2.56$, p value $= 4.5 \times 10^{-12}$) (Table 1, Figures 1 and 2, and Figures S1–S3).

The lead SNP in EA, rs11382794, is located in the intron of cystic fibrosis transmembrane conductance regulator (*CFTR* [MIM: 602421]), the causal gene for cystic fibrosis (CF [MIM: 219700]). CF causes abnormally thick mucus that blocks airways, leading to chronic infections, persistent inflammation, airway remodeling, and progressive respiratory failure.[16,17] Both acute and chronic lung infections are major contributors to morbidity and mortality in individuals with CF.[18,19] In a study of 19,802 CF carriers (individuals with one defective copy of *CFTR*) and 99,010 control individuals, CF carriers were more likely than non-carriers to have pneumonia (OR $= 1.16$), a personal history of recurrent pneumonia (OR $= 2.76$), and other respiratory infections.[20]

The top finding in AA, rs334, is a nonsynonymous variant in hemoglobin subunit beta (*HBB* [MIM:141900]) and the causal mutation for SCD, including sickle cell

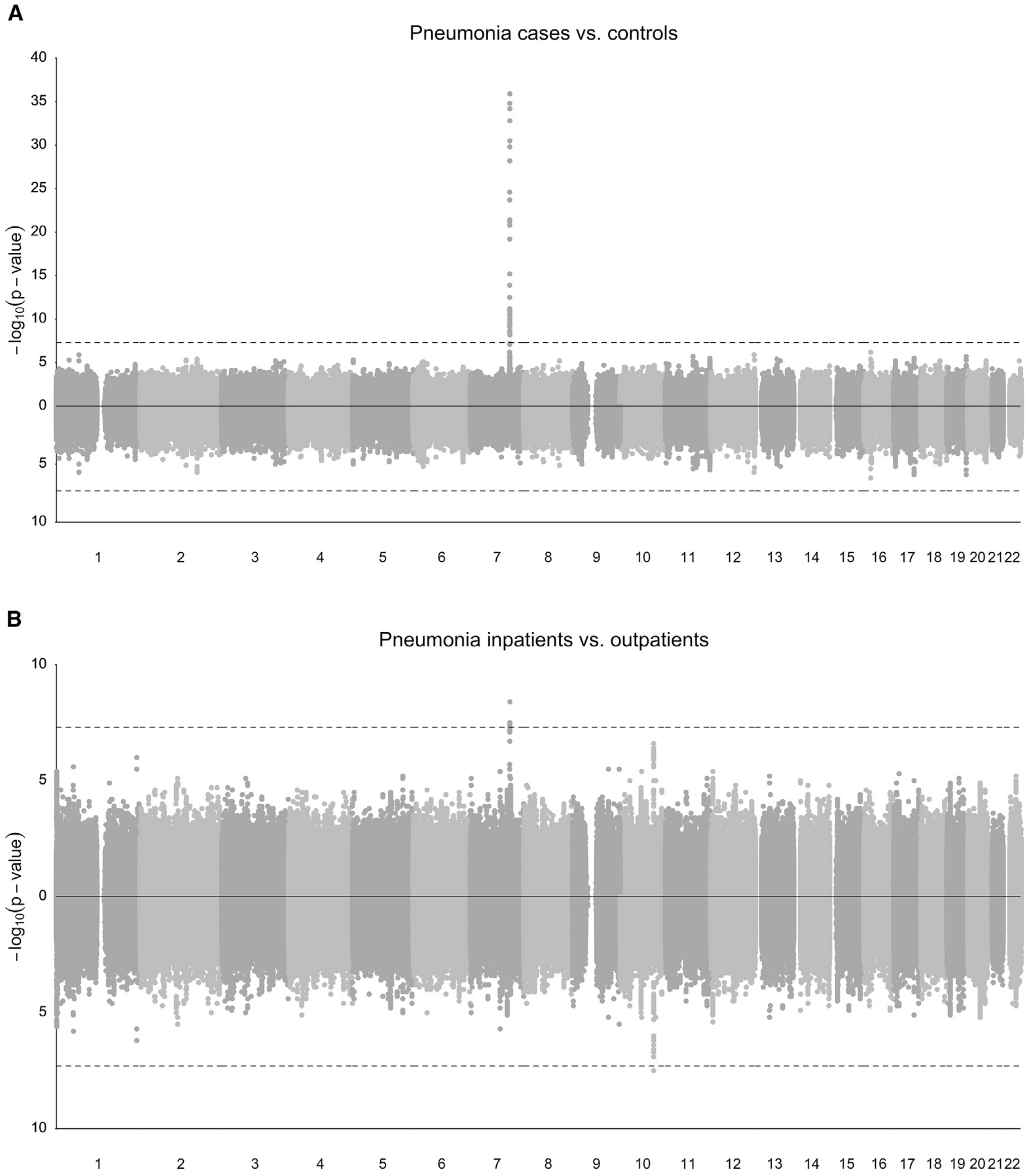


Figure 1. SNP associations with pneumonia in BioVU MEGA^{EX}-genotyped EA
(A) Miami plot of pneumonia susceptibility (individual with pneumonia versus control individuals) with (top) and without (bottom) individuals diagnosed with CF.
(B) Miami plot of pneumonia severity (pneumonia inpatients versus outpatients) with (top) and without (bottom) individuals diagnosed with CF.

anemia (SCA [MIM: 603903]). The relationship between SCA and pneumonia risk has been previously described in epidemiological studies.[21] Children with SCA are more likely to have pneumonia and influenza (OR = 7.38) and acute respiratory infections (OR = 1.29).[22] Acute chest syndrome (ACS) is a common complication of SCD and can be triggered by pneumonia and vaso-occlusive crises and is the leading cause of death in individuals with SCA.[23] Clinically differentiating between ACS and pneumonia can be difficult, and they often overlap in the EHR.

Our two initial findings provide genetic support of previously observed epidemiological predictors of severe pneumonia risk, namely, the two autosomal recessive disorders CF and SCD. Previous studies have also reported associations between carrier status for *CFTR* and CF-associated symptoms.[20] In contrast, SCA carriers had slightly lower risk of pneumonia and influenza (OR = 0.93) compared to subjects with normal hemoglobin,[22] although pneumonia severity was not studied. We applied Fisher's exact test to investigate the effect of the risk allele in

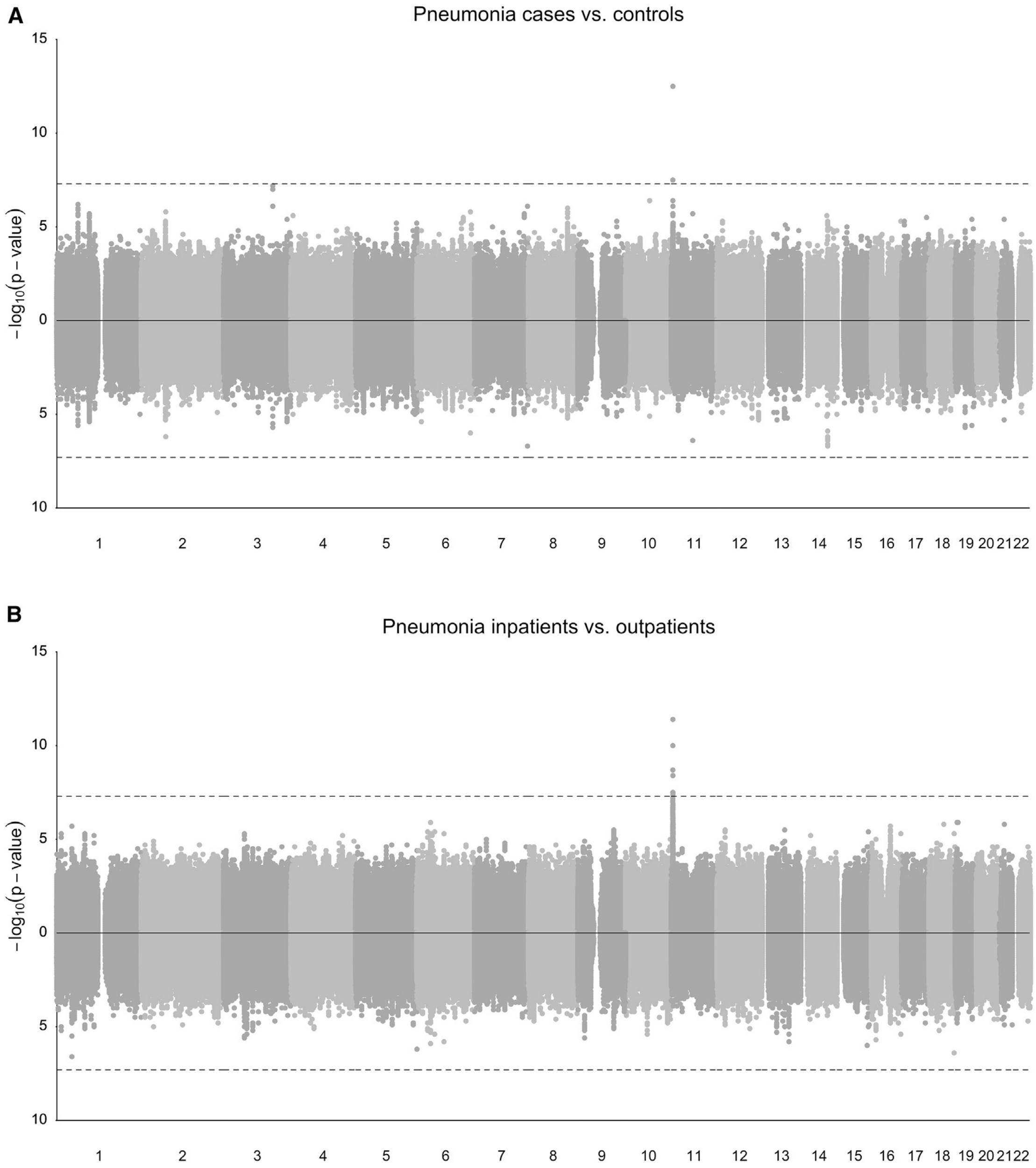


Figure 2. SNP associations with pneumonia in BioVU MEGA^{EX}-genotyped AA
(A) Miami plot of pneumonia susceptibility (individual with pneumonia versus control individuals) with (top) and without (bottom) individuals diagnosed with CF and SCD.
(B) Miami plot of pneumonia severity (pneumonia inpatients versus outpatients) with (top) and without (bottom) individuals diagnosed with CF and SCD.

heterozygous carriers of the risk allele compared to those homozygous for the reference allele. We found heterozygous carriers of rs113827944 (*CFTR*) were at greater risk of developing pneumonia (OR = 1.38, p value = 1.9×10^{-7}), including in sensitivity analyses excluding all individuals with diagnosed CF (OR = 1.17, p value = 0.019; Table 2). Heterozygous carriers of rs334 (*HBB*) were at greater risk of pneumonia-associated hospitalization (OR = 1.55, p value = 0.023), and sensitivity analyses excluding 138 individuals diagnosed with CF and SCD slightly attenuated this effect (OR = 1.49, p value = 0.109), most likely because of a reduction in power. Heterozygote effects observed herein are consistent with an intermediate phenotype or an undetected second variant contributing to compound heterozygosity.[20,24,25]

To identify additional effects masked by the strong effects of *CFTR* and *HBB*, we repeated GWASs after removing individuals diagnosed with CF and SCD. In EA, after removal of individuals with CF, we identified a genome-wide significant signal in *R3HCC1L* (rs10786398, MAF = 30.7%, OR = 1.22, p value = 3.5×10^{-8}, Table 1, and Figures 1 and 3). Replication of effects at this gene were

Table 2. Homozygous and heterozygous effects of CFTR and HBB lead SNPs

										Homozygous[a]		Heterozygous[b]	
Gene	**SNP**	**A1**	**A2**	**Pop**	**Case/control**	**n**	**A1A1**	**A1A2**	**A2A2**	**OR**	**p value**	**OR**	**p value**
CFTR	rs113827944	A	G	EA	pneumonia	8,685	50	339	8,296	23.07	3.7E−32	1.38	1.9E−07
					population controls	59,781	15	1,713	58,053				
					pneumonia, individuals with CF removed	8,504	1	281	8,222	2.34	0.413	1.17	0.019
					population controls, individuals with CF removed	59,695	3	1,694	57,998				
					pneumonia inpatient	5,641	46	236	5,359	6.25	1.7E−05	1.26	0.063
					pneumonia outpatient	3,044	4	103	2,937				
					pneumonia inpatient, individuals with CF removed	5,494	1	187	5,306	N/A	N/A	1.09	0.526
					pneumonia outpatient, individuals with CF removed	3,010	0	94	2,916				
HBB	rs334	A	T	AA	pneumonia	1,617	53	139	1,425	7.43	3.9E−22	1.16	0.124
					population controls	13,214	60	1,021	12,133				
					pneumonia, individuals with CF and SCD removed	1,479	1	84	1,394	2.17	0.421	0.93	0.565
					population controls, individuals with CF and SCD removed	12,822	4	782	12,036				
					pneumonia inpatient	984	48	96	840	6.44	2.1E−06	1.55	0.023
					pneumonia outpatient	633	5	43	585				
					pneumonia inpatient, individuals with CF and SCD removed	875	1	57	817	N/A	N/A	1.49	0.109
					pneumonia outpatient, individuals with CF and SCD removed	604	0	27	577				

[a] Homozygous: A1A1 versus others.
[b] Heterozygous: A1A2 versus A2A2.

observed in a distinct haplotype in BioVU AA (rs7086391, MAF = 16.9%, OR = 1.42, p value = 1.14×10^{-3}, multiple-testing-corrected p value = 0.0490, Figure S4) and at a distinct sentinel variant in UK Biobank EA (rs884811, MAF = 44.2%, OR = 1.05, p value = 3.49×10^{-3}, multiple-testing-corrected p value = 0.0198, Table 1). We note that this region of the genome exhibits low linkage disequilibrium: the lead variants from our discovery analyses and from the UK Biobank validation had an R^2 of 0.55 in the BioVU data despite their being less than 3 kb apart (Figure S5). We also observed effects of rs10786398 in the latest release of the COVID-19 Host Genetics Initiative GWAS comparing hospitalized individuals with COVID-19 to non-hospitalized affected individuals (OR = 1.19, MAF = 31.4%, p value = 0.037),[26] indicating relevance of this variant for susceptibility to and severity of COVID-19. A previously published meta-analysis of lung function in UK Biobank and the SpiroMeta consortium provided additional validation of lead variants in *R3HCC1L* (Table S6).[27] *R3HCC1L* encodes a coiled-coil domain-containing protein. Variants in its promoter region have been reported to be associated with inflammatory skin disease[28] and body mass index.[29,30] However, the molecular function of *R3HCC1L* is still unclear and more studies are needed to clarify its role in the pathogenesis of pneumonia.

In AA, after removal of individuals with CF and SCD, an additional SNP near *UQCRFS1* (MIM: 191327) and non-coding RNA LOC105372352 (rs148218440, MAF = 2.4%; OR = 2.21, p value = 3.6×10^{-8}; Figures S6 and S7) was identified in the comparison of AA pneumonia inpatients and age-, sex-, and ancestry-matched control individuals. *UQCRFS1* encodes a Rieske iron-sulfur protein, which is part of the mitochondrial respiratory chain.[31] Rare mutations in *UQCRFS1* have been previously reported as the cause of a mitochondrial disorder (MIM: 618775).[32] Further, *Uqcrfs1* deletion in mice abolishes mitochondrial reactive oxygen species that are required for antigen specific T cell responses, a hallmark of the adaptive immune response.[33] We did not observe replication of rs148218440 in BioVU non-MEGA^{EX} AA (Table 1), however we note that this analysis was underpowered to detect effect (n = 292 affected individuals and 1,145 control individuals, Table S4); UK Biobank did not have enough

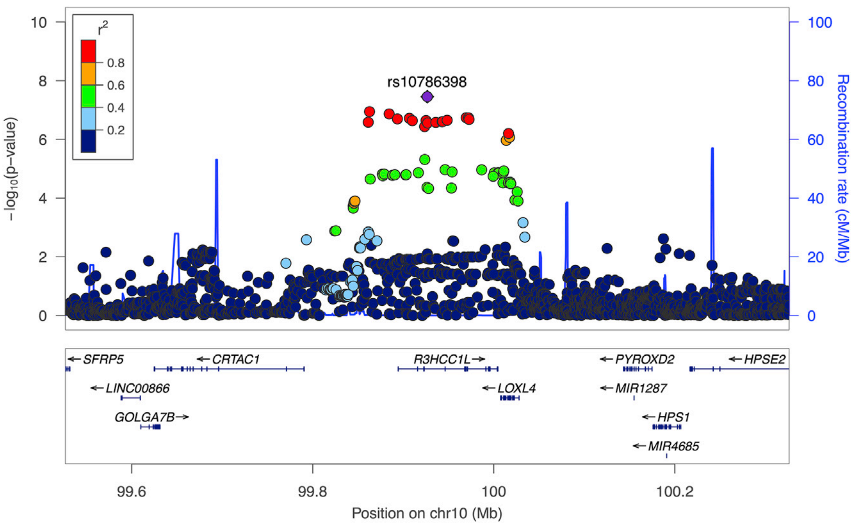


Figure 3. LocusZoom plot of *R3HCC1L* with the genome-wide association study results of pneumonia inpatients versus outpatients from BioVU MEGA^{EX}-genotyped EA excluding CF-affected individuals
Linkage disequilibrium and recombination rate estimates come from 1000 Genomes Project phase 3 European superpopulation reference data.

AA individuals with pneumonia to perform a replication analysis ($n = 76$ affected individuals and 9,272 control individuals, Table S5).

To further investigate the functional role of associated variation at our lead genes, we utilized PrediXcan to impute the genetically regulated expression (GReX) in MEGA^{EX}-genotyped EA and AA with the tissue-specific models trained by GTEx V8.[34,35] The association tests were conducted with SUGEN and adjusted for three principal components, sex, and age (Table S7 and Figures S8 and S9). The GReX of *R3HCC1L* in lungs is significantly associated with severity of pneumonia in EA ($\beta = 1.140$, p value $= 2.0 \times 10^{-7}$). In addition, the association of the GReX of *HBB* with pneumonia susceptibility in EA offers additional evidence of its role ($\beta = 30.670$, p value $= 7.3 \times 10^{-4}$). GReX of *CFTR* is not well captured in lung tissue,[34] however GReX of this gene in other tissues are significantly associated with pneumonia susceptibility in EA (heart left ventricle, $\beta = 0.464$, p value $= 3.7 \times 10^{10}$).

Moreover, we estimated the heritability of pneumonia susceptibility and severity in MEGA^{EX}-genotyped EA. We used genome-wide complex trait analysis (GCTA)[36] to construct a genetic relatedness matrix and calculated the proportion of phenotypic variance explained by the matrix. The SNP-based heritability (h^2) for susceptibility to pneumonia is estimated to be 0.029 and 0.026 after removing individuals diagnosed with CF (Table S8). As expected due to greater similarity of exposure in affected individuals and control individuals, genetic variation explains a larger portion of pneumonia severity compared to susceptibility ($h^2 = 0.121$; after removing individuals with CF, $h^2 = 0.150$). Our results are similar to previous estimates in UK Biobank (0.075 for self-reported pneumonia, 0.15 for pneumonia inpatient).[37]

The data used for the present discoveries exhibit several key limitations. Extracting accurate phenotypes from EHR data is a well-known challenge[38] and most likely results in some misclassification of cases. Such misclassification most likely reduces power to detect effects but is unlikely to introduce systemic bias that would invalidate results. Additionally, distinguishing viral and bacterial pneumonia via clinical criteria is challenging, and data on positive cultures of blood or pleural fluid for bacterial or positive nasopharyngeal sample or positive serology for viral were not widely available in BioVU. Despite these challenges, we were able to detect four loci linked to pneumonia, including two regions. Further research will be required to differentiate the genetic architecture of specific pneumonia etiologies and to confirm the *UQCRFS1* association, and functional studies will be needed to determine the biological mechanism underlying both signals.

In summary, we leveraged a large-scale biobank to identify genetic variants associated with the susceptibility and severity of pneumonia. Two clinically relevant Mendelian disease genes, *CFTR* and *HBB*, were implicated. These important genetic results indicate that individuals with CF and SCA are at heightened risk for development of and severe outcomes from pneumonia, an effect which may translate to COVID-19 outcomes. Heterozygous carriers of the CF risk allele demonstrated elevated risk of pneumonia susceptibility, and carriers of the SCD risk allele demonstrated elevated risk for pneumonia severity. These findings may have important implications for genetically informed patient care in infectious lung disease. They are also critically important in the context of the COVID-19 pandemic, and future studies will be needed to establish whether these carriers exhibit a silent, heightened risk for poor outcomes from COVID-19 as well. We also identified two additional pneumonia-related genes: *R3HCC1L* and *UQCRFS1*. Although we were most likely underpowered to replicate effects of *UQCRFS1* variation in AA in our data resources, we successfully replicated *R3HCC1L* effects in two independent datasets and validated effects of *R3HCC1L* in independent GWASs of COVID-19[26] and lung function.[27] Characterizing host genome effects on infectious disease susceptibility and severity can offer important insight into the molecular etiology of risk; our findings may help elucidate pathophysiological processes for pneumonia, an important COVID-19 sequelae.

Data and code availability

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Supplemental Data

Supplemental Data can be found online at https://doi.org/10.1016/j.ajhg.2020.12.010.

Acknowledgments

The authors would like to thank Austin McPhilamy for his help identifying ICD codes relevant to pneumonia phenotyping. This project was supported by CTSA award UL1 TR002243 from the National Center for Advancing Translational Sciences. H.-H.C., D.M.S., L.E.P., R.J.B., C.D.H., and J.E.B. were supported by funding from the National Institute of General Medical Sciences, National Institutes of Health, R01GM133169. D.M.S. was supported by funding from the National Institutes of Health, 5T32GM080178.

Declaration of interests

J.A.B. receives support from Omniox for an ARDS and pneumonia project. The other authors declare no competing interests.

Received: August 11, 2020
Accepted: December 2, 2020
Published: December 13, 2020

Web resources

Coronavirus Disease 2019, https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/index.html
COVID-19 Map, https://coronavirus.jhu.edu/map.html
Online Mendelian Inheritance in Man, https://omim.org

References

1. Guo, Y.R., Cao, Q.D., Hong, Z.S., Tan, Y.Y., Chen, S.D., Jin, H.J., Tan, K.S., Wang, D.Y., and Yan, Y. (2020). The origin, transmission and clinical therapies on coronavirus disease 2019 (COVID-19) outbreak - an update on the status. Mil. Med. Res. *7*, 11.
2. Torres, A., Blasi, F., Dartois, N., and Akova, M. (2015). Which individuals are at increased risk of pneumococcal disease and why? Impact of COPD, asthma, smoking, diabetes, and/or chronic heart disease on community-acquired pneumonia and invasive pneumococcal disease. Thorax *70*, 984–989.
3. Hussain, A., Bhowmik, B., and do Vale Moreira, N.C. (2020). COVID-19 and diabetes: Knowledge in progress. Diabetes Res. Clin. Pract. *162*, 108142.
4. Chhiba, K.D., Patel, G.B., Vu, T.H.T., Chen, M.M., Guo, A., Kudlaty, E., Mai, Q., Yeh, C., Muhammad, L.N., Harris, K.E., et al. (2020). Prevalence and characterization of asthma in hospitalized and non-hospitalized patients with COVID-19. J. Allergy Clin. Immunol. *146*, 307–314.e4.
5. Zhao, Q., Meng, M., Kumar, R., Wu, Y., Huang, J., Lian, N., Deng, Y., and Lin, S. (2020). The impact of COPD and smoking history on the severity of COVID-19: A systemic review and meta-analysis. J. Med. Virol. *92*, 1915–1921.
6. Dietz, W., and Santos-Burgoa, C. (2020). Obesity and its Implications for COVID-19 Mortality. Obesity (Silver Spring) *28*, 1005.
7. Rautanen, A., Mills, T.C., Gordon, A.C., Hutton, P., Steffens, M., Nuamah, R., Chiche, J.D., Parks, T., Chapman, S.J., Davenport, E.E., et al.; ESICM/ECCRN GenOSept Investigators (2015). Genome-wide association study of survival from sepsis due to pneumonia: an observational cohort study. Lancet Respir. Med. *3*, 53–60.
8. Zúñiga, J., Buendía-Roldán, I., Zhao, Y., Jiménez, L., Torres, D., Romo, J., Ramírez, G., Cruz, A., Vargas-Alarcon, G., Sheu, C.C., et al. (2012). Genetic variants associated with severe pneumonia in A/H1N1 influenza infection. Eur. Respir. J. *39*, 604–610.
9. Hayden, L.P., Cho, M.H., McDonald, M.N., Crapo, J.D., Beaty, T.H., Silverman, E.K., Hersh, C.P.; and COPDGene Investigators * (2017). Susceptibility to Childhood Pneumonia: A Genome-Wide Analysis. Am. J. Respir. Cell Mol. Biol. *56*, 20–28.
10. Tian, C., Hromatka, B.S., Kiefer, A.K., Eriksson, N., Noble, S.M., Tung, J.Y., and Hinds, D.A. (2017). Genome-wide association and HLA region fine-mapping studies identify susceptibility loci for multiple common infections. Nat. Commun. *8*, 599.
11. Campos, A., Kho, P.F., Vazquez-Prada, K.X., Garcia-Marin, L.M., Martin, N.G., Cuellar-Partida, G., and Renteria, M.E. (2020). Genetic susceptibility to pneumonia: A GWAS meta-analysis between UK Biobank and FinnGen. medRxiv. https://doi.org/10.1101/2020.06.22.20103556.
12. Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A.E., Kwong, A., Vrieze, S.I., Chew, E.Y., Levy, S., McGue, M., et al. (2016). Next-generation genotype imputation service and methods. Nat. Genet. *48*, 1284–1287.
13. Lin, D.Y., Tao, R., Kalsbeek, W.D., Zeng, D., Gonzalez, F., 2nd, Fernández-Rhodes, L., Graff, M., Koch, G.G., North, K.E., and Heiss, G. (2014). Genetic association analysis under complex survey sampling: the Hispanic Community Health Study/Study of Latinos. Am. J. Hum. Genet. *95*, 675–688.
14. Staples, J., Qiao, D., Cho, M.H., Silverman, E.K., Nickerson, D.A., Below, J.E.; and University of Washington Center for Mendelian Genomics (2014). PRIMUS: rapid reconstruction of pedigrees from genome-wide estimates of identity by descent. Am. J. Hum. Genet. *95*, 553–564.
15. Li, H., Glusman, G., Hu, H., Shankaracharya, Caballero, J., Hubley, R., Witherspoon, D., Guthery, S.L., Mauldin, D.E., Jorde, L.B., et al. (2014). Relationship estimation from whole-genome sequence data. PLoS Genet. *10*, e1004144.
16. Ratjen, F., and Döring, G. (2003). Cystic fibrosis. Lancet *361*, 681–689.
17. Rowe, S.M., Miller, S., and Sorscher, E.J. (2005). Cystic fibrosis. N. Engl. J. Med. *352*, 1992–2001.
18. Davis, P.B. (2006). Cystic fibrosis since 1938. Am. J. Respir. Crit. Care Med. *173*, 475–482.
19. Sanders, D.B., and Fink, A.K. (2016). Background and Epidemiology. Pediatr. Clin. North Am. *63*, 567–584.
20. Miller, A.C., Comellas, A.P., Hornick, D.B., Stoltz, D.A., Cavanaugh, J.E., Gerke, A.K., Welsh, M.J., Zabner, J., and Polgreen, P.M. (2020). Cystic fibrosis carriers are at increased risk for a wide range of cystic fibrosis-related conditions. Proc. Natl. Acad. Sci. USA *117*, 1621–1627.
21. Poehling, K.A., Light, L.S., Rhodes, M., Snively, B.M., Halasa, N.B., Mitchel, E., Schaffner, W., Craig, A.S., and Griffin, M.R.

(2010). Sickle cell trait, hemoglobin C trait, and invasive pneumococcal disease. Epidemiology *21*, 340–346.

22. Reeves, S.L., Jary, H.K., Gondhi, J.P., Kleyn, M., and Dombkowski, K.J. (2019). Health outcomes and services in children with sickle cell trait, sickle cell anemia, and normal hemoglobin. Blood Adv. *3*, 1574–1580.
23. Jain, S., Bakshi, N., and Krishnamurti, L. (2017). Acute Chest Syndrome in Children with Sickle Cell Disease. Pediatr. Allergy Immunol. Pulmonol. *30*, 191–201.
24. Agarwal, S., Tamhankar, P.M., Kumar, R., and Dalal, A. (2010). Clinical and haematological features in a compound heterozygote (HBB:c.92 + 5G > C/HBB:c.93-2A > C) case of thalassaemia major. Int. J. Lab. Hematol. *32*, 369–372.
25. Raju, S.V., Tate, J.H., Peacock, S.K., Fang, P., Oster, R.A., Dransfield, M.T., and Rowe, S.M. (2014). Impact of heterozygote CFTR mutations in COPD patients with chronic bronchitis. Respir. Res. *15*, 18.
26. Covid-19 Host Genetics Initiative (2020). The COVID-19 Host Genetics Initiative, a global initiative to elucidate the role of host genetic factors in susceptibility and severity of the SARS-CoV-2 virus pandemic. Eur. J. Hum. Genet. *28*, 715–718.
27. Shrine, N., Guyatt, A.L., Erzurumluoglu, A.M., Jackson, V.E., Hobbs, B.D., Melbourne, C.A., Batini, C., Fawcett, K.A., Song, K., Sakornsakolpat, P., et al.; Understanding Society Scientific Group (2019). New genetic signals for lung function highlight pathways and chronic obstructive pulmonary disease associations across multiple ancestries. Nat. Genet. *51*, 481–493.
28. Baurecht, H., Hotze, M., Brand, S., Büning, C., Cormican, P., Corvin, A., Ellinghaus, D., Ellinghaus, E., Esparza-Gordillo, J., Fölster-Holst, R., et al.; Psoriasis Association Genetics Extension (2015). Genome-wide comparative analysis of atopic dermatitis and psoriasis gives insight into opposing genetic mechanisms. Am. J. Hum. Genet. *96*, 104–120.
29. Hoffmann, T.J., Choquet, H., Yin, J., Banda, Y., Kvale, M.N., Glymour, M., Schaefer, C., Risch, N., and Jorgenson, E. (2018). A Large Multiethnic Genome-Wide Association Study of Adult Body Mass Index Identifies Novel Loci. Genetics *210*, 499–515.
30. Zhu, Z., Guo, Y., Shi, H., Liu, C.L., Panganiban, R.A., Chung, W., O'Connor, L.J., Himes, B.E., Gazal, S., Hasegawa, K., et al. (2020). Shared genetic and experimental links between obesity-related traits and asthma subtypes in UK Biobank. J. Allergy Clin. Immunol. *145*, 537–549.
31. Sánchez, E., Lobo, T., Fox, J.L., Zeviani, M., Winge, D.R., and Fernández-Vizarra, E. (2013). LYRM7/MZM1L is a UQCRFS1 chaperone involved in the last steps of mitochondrial Complex III assembly in human cells. Biochim. Biophys. Acta *1827*, 285–293.
32. Gusic, M., Schottmann, G., Feichtinger, R.G., Du, C., Scholz, C., Wagner, M., Mayr, J.A., Lee, C.Y., Yépez, V.A., Lorenz, N., et al. (2020). Bi-Allelic UQCRFS1 Variants Are Associated with Mitochondrial Complex III Deficiency, Cardiomyopathy, and Alopecia Totalis. Am. J. Hum. Genet. *106*, 102–111.
33. Sena, L.A., Li, S., Jairaman, A., Prakriya, M., Ezponda, T., Hildeman, D.A., Wang, C.R., Schumacker, P.T., Licht, J.D., Perlman, H., et al. (2013). Mitochondria are required for antigen-specific T cell activation through reactive oxygen species signaling. Immunity *38*, 225–236.
34. Barbeira, A.N., Bonazzola, R., Gamazon, E.R., Liang, Y., Park, Y., Kim-Hellmuth, S., Wang, G., Jiang, Z., Zhou, D., Hormozdiari, F., et al. (2019). Widespread dose-dependent effects of RNA expression and splicing on complex diseases and traits. bioRxiv. https://doi.org/10.1101/814350.
35. Aguet, F., Barbeira, A.N., Bonazzola, R., Brown, A., Castel, S.E., Jo, B., Kasela, S., Kim-Hellmuth, S., Liang, Y., Oliva, M., et al. (2019). The GTEx Consortium atlas of genetic regulatory effects across human tissues. bioRxiv. https://doi.org/10.1101/787903.
36. Yang, J., Lee, S.H., Goddard, M.E., and Visscher, P.M. (2011). GCTA: a tool for genome-wide complex trait analysis. Am. J. Hum. Genet. *88*, 76–82.
37. Ge, T., Chen, C.Y., Neale, B.M., Sabuncu, M.R., and Smoller, J.W. (2018). Correction: Phenome-wide heritability analysis of the UK Biobank. PLoS Genet. *14*, e1007228.
38. DeLisle, S., Kim, B., Deepak, J., Siddiqui, T., Gundlapalli, A., Samore, M., and D'Avolio, L. (2013). Using the electronic medical record to identify community-acquired pneumonia: toward a replicable automated strategy. PLoS ONE *8*, e70944.

International Journal of
Molecular Sciences

Review

Amoebiasis: Advances in Diagnosis, Treatment, Immunology Features and the Interaction with the Intestinal Ecosystem

Patricia Morán [1], Angélica Serrano-Vázquez [1], Liliana Rojas-Velázquez [1], Enrique González [1], Horacio Pérez-Juárez [1], Eric G. Hernández [1], Maria de los Angeles Padilla [1], Martha E. Zaragoza [1], Tobías Portillo-Bobadilla [2], Manuel Ramiro [3] and Cecilia Ximénez [1,*]

[1] Laboratorio de Inmunología, Unidad de Investigación en Medicina Experimental, Facultad de Medicina, Universidad Nacional Autónoma de México (UNAM), Ciudad de México 06726, Mexico
[2] Unidad de Bioinformática, Bioestadística y Biología Computacional, Red de Apoyo a la Investigación, Coordinación de la Investigación Científica, Universidad Nacional Autónoma de México (UNAM)-Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán, Ciudad de México 14080, Mexico
[3] División de Estudios de Posgrado, Facultad de Medicina, Universidad Nacional Autónoma de México (UNAM), Ciudad de México 04510, Mexico
[*] Correspondence: cximenez@unam.mx

Abstract: This review of human amoebiasis is based on the most current knowledge of pathogenesis, diagnosis, treatment, and Entamoeba/microbiota interactions. The most relevant findings during this last decade about the *Entamoeba* parasite and the disease are related to the possibility of culturing trophozoites of different isolates from infected individuals that allowed the characterization of the multiple pathogenic mechanisms of the parasite and the understanding of the host–parasite relationship in the human. Second, the considerable advances in molecular biology and genetics help us to analyze the genome of *Entamoeba*, their genetic diversity, and the association of specific genotypes with the different amoebic forms of human amoebiasis. Based on this knowledge, culture and/or molecular diagnostic strategies are now available to determine the *Entamoeba* species and genotype responsible for invasive intestinal or extraintestinal amoebiasis cases. Likewise, the extensive knowledge of the immune response in amoebiasis with the appearance of new technologies made it possible to design diagnostic tools now available worldwide. Finally, the understanding of the interaction between the *Entamoeba* species and the intestinal microbiota aids the understanding of the ecology of this parasite in the human environment. These relevant findings will be discussed in this review.

Keywords: amoebiasis; advanced diagnosis and treatment; pathogenesis; intestinal microbiota interaction

Citation: Morán, P.; Serrano-Vázquez, A.; Rojas-Velázquez, L.; González, E.; Pérez-Juárez, H.; Hernández, E.G.; Padilla, M.d.l.A.; Zaragoza, M.E.; Portillo-Bobadilla, T.; Ramiro, M.; et al. Amoebiasis: Advances in Diagnosis, Treatment, Immunology Features and the Interaction with the Intestinal Ecosystem. *Int. J. Mol. Sci.* **2023**, *24*, 11755. https://doi.org/10.3390/ijms241411755

Academic Editors: Leszek Rolbiecki and Joanna N. Izdebska

Received: 27 May 2023
Revised: 12 July 2023
Accepted: 19 July 2023
Published: 21 July 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The genus *Entamoeba* consists of at least seven different species (*E. histolytica*, *E. coli*, *E. hartmanni*, *E. polecki*, *E. dispar*, *E. moshkovskii*, and *E. bangladeshi*) that can inhabit the human intestine and one (*E. gingivalis*) that can be found in the human oral cavity [1]. Until today, the only species recognized as an etiologic agent of the amoebic invasive disease is *Entamoeba histolytica*, with the large intestine as the main target organ [2].

Amoebiasis is defined as the gastrointestinal infection caused by *E. histolytica* with or without clinical manifestations [2], and is responsible for the worldwide morbidity and mortality rates due to amoebiasis in children and young adults. The global significance of amoebiasis is widespread, with the highest burden of amoebiasis borne by those residing in developing countries, particularly the tropics and subtropics, where there are inadequate hygiene levels and access to sanitation [3]. The efforts being made in the study of amoebiasis epidemiology in developing countries allow us to have a deeper understanding of its global panorama [4,5]. The estimations of the worldwide burden of amoebiasis by the World Health Organization (WHO) indicated that approximately 500 million people were infected with the parasite and 10% of these individuals had invasive amoebiasis [6]. Some reports mention that more than 55,000 people die each year due to acute intestinal amoebiasis [7,8].

Amoebiasis is caused by the ingestion of food or water contaminated with cysts, the infective form of *E. histolytica*. Following excystation, the emerging trophozoites migrate to the large intestine [9]. The trophozoites (mainly for *E. histolytica*) can become virulent and invasive, where they start to destroy the mucous–epithelial barrier, thus inducing the overproduction of mucus, killing host cells, and provoking inflammation, subsequently causing amoebic colitis [10]. *Entamoeba histolytica* can be transmitted by asymptomatic hosts infected with cysts and patients with intestinal amoebiasis, who are the natural transmitters [11]. The *Entamoeba* cysts are resistant to desiccation in soil and can survive in humid environments and food or water for several weeks, and susceptible hosts exposed to the infection sources ingest the cysts, which then undergo excystation during their pass through the gastrointestinal tract [11,12]. When in excystation, each cyst produces eight vegetative forms or trophozoites, which are the motile form of the parasite [11]. Trophozoites multiply by binary fission and some of them may encyst and be excreted with stools [13]. In asymptomatic infections, *Entamoeba* species trophozoites live as commensal feeding on colonic microbiota and nutrients of the host and forming cysts that pass through stools to perpetuate a new life cycle by fecal–oral spreading (Figure 1) [4]. It should be noted that more than 90% of infections have an asymptomatic course and are often self-limited over different periods [14].

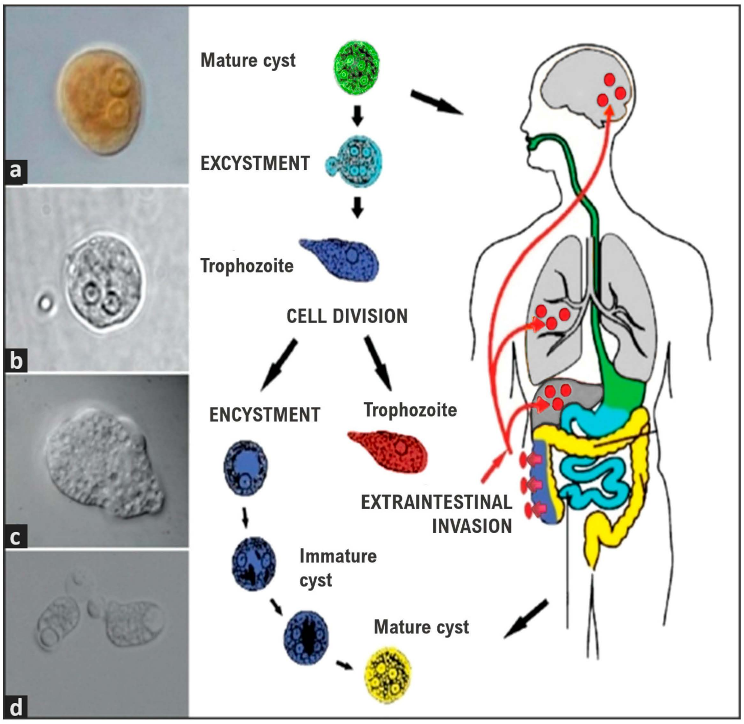


Figure 1. The life cycle of *E. histolytica*/*E. dispar.* Mature cysts (green) are swallowed along with food and water. The cysts pass through the stomach, but if the stomach is alkaline or when the cyst reaches the small intestine the cyst wall is damaged by trypsin, excystation starts (light blue). Trophozoites emerge (blue) and colonize the caecum and colon. Trophozoites are the infective amoeboid form which multiply by binary fission and can migrate to other tissues causing an extraintestinal invasion (red) or initiate encystment forming immature cysts and, finally, a mature cyst (yellow) to start a new cycle. (**a**) Mature cyst stained with 4% Lugol's solution (100× magnification); (**b**) Mature cyst without staining (100×); (**c**) Trophozoite observed with differential interference contrast (DIC) (100×); (**d**) Trophozoites of *E. histolytica* with phagocyted erythrocytes (DIC) (40×) in Ximénez et al. [11].

Amoebiasis is endemic in the developing countries of Central and South America, Africa, and Asia [3,15]. However, until now, the exact burden of amoebiasis in these countries has been difficult to determine, and reports can be affected by geographic region, study design, sample size, incubation, symptom severity, and the sensitivity of the diagnostic modality used. In addition, diagnostic capacities and surveillance are often limited in areas where *E. histolytica* is endemic [4,16].

2. Clinical Features

Following the 1997 documents by the World Health Organization, amoebiasis is defined as the infection of the protozoan parasite *E. histolytica* with or without clinical manifestations [2]. The spectrum of the disease ranges from asymptomatic infection to the development of a severe infection with amoebic colitis and/or amoebic liver abscesses [17]. Chronic non-dysenteric colitis is the most frequent form of amoebiasis in people of all ages, characterized by non-specific symptoms [3]. After exposure to the parasite, 80% of patients display symptoms over a few days to 4–6 weeks [11]. The natural history of invasive intestinal amoebiasis is an acute event, characterized by the presence of diarrhea that occurs days or weeks after exposure and does not last more than four to five weeks, although there are reports of its occurrence years later after exposure to a source of infection [11]. In the latter case, it is assumed that the cause–effect relationship is extremely difficult to corroborate [18]. The natural history of this clinical form can also mimic irritable bowel syndrome. Symptoms more relevant in this instance are periods of abdominal pain (colic) and auto-limited episodes of diarrhea alternating with constipation. However, both non-dysenteric amoebic colitis and irritable bowel syndrome are controversial themes in clinical practice [11]. A few individuals with *E. histolytica* infections develop extraintestinal features including liver abscesses, purulent pericarditis, lung abscesses, pneumonia, peritonitis skin lesions, and even cerebral abscesses, due to the dissemination of trophozoites via the blood circulatory system [19–21]. Of these clinical features, the liver abscess is the most common manifestation, and delay in diagnosis and treatment may cause fatalities.

2.1. Amoebic Liver Abscess

Amoebic liver abscess is a disease that can affect individuals of all ages; in some endemic areas, the incidence rates are higher in both children under 5 years and young adults. Males are also more prone to developing amoebic hepatic abscesses than females [3,11]. The most common symptoms suggestive of amoebic liver abscess are fever (38 °C), chills and diaphoresis, anorexia, and abdominal pain in the right upper quadrant that increases during inspiration. Pain also frequently radiates to the shoulder and back. Nausea has also been referenced but diarrhea is only occasionally mentioned (50% of cases) [11,22].

Hepatomegaly can be detected during digital percussion of the hepatic area and is always related to the dimensions of the abscess; patients can also display peritoneal signs (abdominal guarding or rebound). The absence of intestinal noises, jaundice, and pleural or pericardial rub are symptoms that should elicit alarm related to the rupture or imminence of rupture of the hepatic abscess [23].

2.2. The Severe Forms of Amoebiasis

The severe forms of amoebiasis are colon ameboma, fulminant necrotizing colitis, and toxic megacolon [24]. The appearance of symptoms, such as severe dysentery and pain with signs of peritoneal irritation (rebound), intense tenesmus, a fever (>38 °C), tachycardia, hypertension, nausea, and anorexia, are suggestive of the previously mentioned severe forms of intestinal amoebiasis [20]. The mortality rates of dysenteric syndrome due to *E. histolytica* are less than 1%, but mortality due to complications increases up to 75% [3]. Fulminant amoebic colitis is a rare complication of amoebiasis associated with high mortality, and can occur in more than 50% of cases with severe colitis [20,24]. The severe forms of invasive amoebiasis can be observed in young children, pregnant women, the elderly, and those with chronic diseases and individuals being treated with immunosuppressants or those with immunodeficiency disorders [11].

The circumstances that define the extraintestinal invasive behavior of some *E. histolytica* strains remain unknown today. However, this seems to be an infrequent event as suggested by the low morbidity rates of amoebic liver abscess and other extraintestinal forms of invasive amoebiasis compared with the prevalence rates of asymptomatic infections and intestinal disease, and fortunately, in the last decade, such complications are uncommon [25,26].

3. Diagnosis

Amoebiasis is an acute disease in which the most frequent symptoms are abdominal pain (colic) and the presence of diarrhea with mucus and/or blood, or a clear dysenteric syndrome, but fever and other systemic symptoms are infrequent [27]. In endemic areas, patients with clinical signs and symptoms including gastrointestinal upset and watery or bloody diarrhea should be suspected of intestinal amoebiasis and should be treated [28].

3.1. Macroscopy and Microscopy Analyses

The gold standard for the diagnosis of amoebiasis is the identification of the parasite, mainly in the feces of people infected with *Entamoeba*, where cysts and, rarely, trophozoites are identified. This identification is carried out by direct microscopy examination; the Faust–Ferreira concentration method has been the most widely used for microscopic identification [29]. While technically simple in approach, these techniques require the expertise of highly qualified technicians in the morphological identification of ova, cysts, and trophozoites to be feasible, and the sensitivity and specificity are no more than 80% [30]. This technique cannot differentiate between *E. histolytica* or *E. dispar*, but in some endemic communities, this is the only diagnostic technology available [31]. Where endoscopy examination is available, a colonoscopy can be of great help in the clinical diagnosis of invasive intestinal amoebiasis [20]. This procedure allows for the microscopic examination of samples taken directly from the characteristic flask-shaped ulcer produced by *E. histolytica* and from other sites of mucosal lesions.

On the other hand, a colonoscopy detects the presence of lesions related to the mentioned severe forms of intestinal amoebiasis and allows for the differential diagnosis of other pathologies, such as inflammatory bowel disease or colon carcinoma (Figure 2) [11].

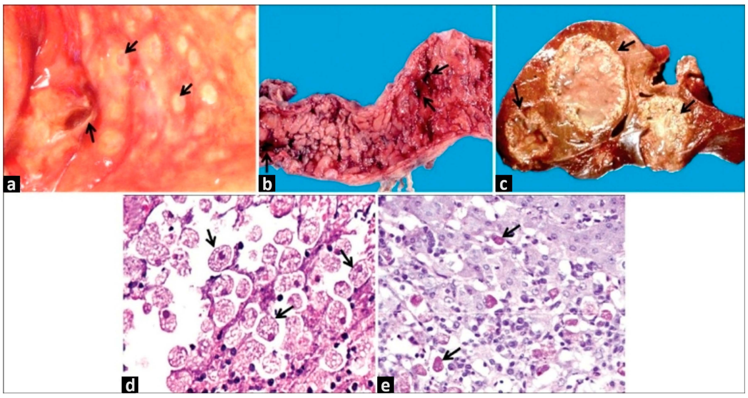


Figure 2. (**a**) Intestinal flask-shaped ulcers observed through rectosigmoidoscopy examination. Arrows indicate colonic ulcers. (**b**) Large bowel necropsy specimen from a case of fulminant amoebic colitis. Arrows indicate hemorrhagic ulcers and important intestinal mucosa necrosis. (**c**) Necropsy specimen of liver abscesses. Arrows indicate the three large abscesses. (**d**) Intestinal biopsy obtained from the edge of a flask-shaped ulcer where large numbers of trophozoites (HE and PAS stained, 60×) are visible. (**e**) Biopsy obtained from the edge of an amoebic liver abscess (HE and PAS stained, 20×). Notice the presence of trophozoites, hepatocytes, and many inflammatory cells. Courtesy Doctor Ruy Pérez-Tamayo in Ximénez et al. [11].

3.2. Imagen Diagnosis

In amoebic liver abscesses, in general, the right hepatic lobule is the most frequently affected due to the portal circulatory system of the right colon; however, the left lobule can be also affected [20]. Laboratory findings suggestive of amoebic liver abscess are the presence of leukocytosis, neutrophilia, increased globular sedimentation velocity, and high levels of alkaline phosphatase [3,32].

Thoracic X-ray data useful in the diagnosis of amoebic abscesses, as well as other types of hepatic abscesses, include the elevation of the right hemidiaphragm, atelectasis, or pleural effusion [3,33]. An ultrasound is the gold standard technique for the diagnosis of amoebic liver abscesses and is of tremendous importance in countries with limited medical and economic resources [34]. Several authors have mentioned the presence of a large single abscess as a frequent characteristic of amoebic liver abscesses; however, this characteristic is not a determinant of amoebic abscesses, and the pyogenic abscesses are also not characteristic of multiple abscesses [35,36].

Early lesions due to the amoebic invasion of hepatic parenchyma are multifocal (micro-abscesses) because of tissue destruction and necrosis by proteases from *E. histolytica*, and neutrophil recruitment to the site of infection [37]. The advantage of computed tomography (CT) scans and magnetic resonance is the detection of small abscesses and the high definition of the images. Moreover, other techniques not always available in endemic countries (e.g., gallium scans) can help differentiate between amoebic (cold images) and pyogenic abscesses (hot images). Thus, the difference is based on the absence (amoebic) or presence (pyogenic) of white blood cells in the abscesses [36,38].

3.3. Immunological and Molecular Diagnosis

During the last 10 years, diagnostics in amoebiasis have changed dramatically, considerably improving sensitivity and specificity. Some of the current techniques are based on immunological strategies, such as ELISA and different modalities of polymerase chain reaction (PCR). There are clear advantages in the bedside diagnosis and clinical laboratories of health institutions [31].

In our experience, immunologically based diagnostics tests for the detection of anti-amoebic antibodies in serum are excellent tests in hospitals, particularly in patients with liver abscesses, and their results are reproducible and reliable to perform the differential diagnosis between amoebic and pyogenic liver abscesses and determine the appropriate treatment [39,40]. The presence of the parasite can also be identified in the drainage of liver abscesses, even though trophozoites are not detected in these types of samples, it is possible to obtain the DNA of the parasite through molecular techniques. Likewise, in tissue biopsies of cutaneous or intestinal amoebiasis lesions, microscopic identification of the trophozoites and/or by PCR is possible. Advances in molecular diagnostics have improved our understanding and led to the recognition and separation of *E. histolytica* from nonpathogenic species of *Entamoeba*. Therefore, PCR has become a very sensitive and specific diagnostic method [31,40].

ELISA tests have been used both for the detection of antiamoebic antibodies in serum and for the capture of amoebic antigens in feces. However, the difficulty of differentiating a past infection from an active one considerably limited its use to identify patients requiring treatment; instead, it has been used for seroepidemiological studies [41–43], so there are many commercial tests that can even differentiate *E. histolytica* from *E. dispar*. These diagnostic kits allow for the generation of reliable results quickly by following simple instructions, but they are expensive due to the use of monoclonal Abs (mAbs) for antigen capture and are not routinely used in clinical laboratories.

Of these kits, the most popular mAbs for use against Gal/Lectin GalNAc are: the *E. histolytica* TechLab kit and the CELISA PATH kit (Cellabs, Brookvale, Australia), which were developed specifically to detect *E. histolytica* in stool samples, and the TechLab *E. histolytica* II (Blacksburg, VA, USA), which allows the detection of different epitopes between the Gal-lectins of *E. histolytica* and *E. dispar* [44–46]. The sensitivity and specificity of the

TechLab kits I and II vary by 80–99% and 86–98%, respectively [44]. Other kits include the Optimum S kit (Merlin Diagnostika, Bornheim-Hersel, Germany), which detects the components of the family of serine-rich proteins, the Alexon ProSpecT ELISA (Remel Inc., Lenexa, KS, USA), which detects an "AND *E. histolytica*-specific antigen" (EHSA) from *E. histolytica*/*E. dispar*, and the Triage Parasite Panel (BIOSITE Diagnostics, San Diego, CA, USA), which detects a 29-kDa antigen from *E. histolytica*/*E. dispar* using an immune-chromatographic strip (118). Recently, an ELISA sandwich assay was developed using two mAbs raised against a spacer region of the Jacob2 lectin of *E. histolytica*, a protein of the cyst wall [47]. Notably, the mAbs only recognized the cysts of *E. histolytica* but did not recognize the cysts of *E. dispar* or *E. Bangladeshi* by immunofluorescence assays. The successful detection of cysts in formalin-fixed stool samples was observed, suggesting that the anti-Jacob2 mAbs could be used for the development of a diagnostic kit for *E. histolytica* infection in any type of stool sample.

Recently, several molecular diagnostic methods, including conventional and real-time polymerase chain reactions (PCRs), have been developed for the specific diagnosis of each of these three *Entamoeba* species in clinical laboratories. Currently, there is a great variety of PCR methods targeting different genes for the detection and differentiation of these amoebas [31,48]. The consistent genetic diversity detected in the small rRNA subunit (18SDNAr) of *E. histolytica* and *E. dispar* initiated the use of this gene as a target for the differentiation of the two species [49–53]. PCR techniques targeting this gene are 100 times more sensitive than the current best available enzyme immunoassay (ELISA) kit [54–56] and are widely used to distinguish these species since there are multiple copies of these genes in amoebas [57], which makes the 18SDNAr gene easier to detect than a DNA fragment from a single copy of a gene.

However, for a good result from the PCR, good-quality DNA is required, which is why extraction optimization is necessary. Recently, simple and effective methods for the isolation of *Entamoeba* DNA from feces have been developed. Commercially available DNA extraction kits are recommended due to their ease and success as they minimize extraction time, and the DNA can be extracted directly from the feces without the need to culture the parasite. The following kits have been used for the direct extraction of DNA from fecal samples include XTRAX DNA extraction kit [58] (Gull Laboratories, Salt Lake City, UT, USA), Extract Master Faecal DNA extraction kit Epicentre Biotechnologies, Madison, WI, USA), and the Genomic DNA Prep Plus kit [59]. Of the methods for DNA extraction, Qiagen kits have been used the most widely with the QIAamp stool kit now predominating. The QIAamp tissue kit spin columns (QIAGEN) have also been used for the extraction and purification of *Entamoeba* DNA from fecal samples [60]. Various modifications, including optimizing the duration and temperature of proteinase K digestion, adding additional wash steps, and treatment of the sample with 2% polyvinylpolypyrrolidone, have all been used [61,62]. The QIAamp stool kit has proven to be the most widely accepted method for DNA extraction providing successful and reliable recovery of DNA from fecal material [50,63,64]. Aspirated pus from ALA can also be tested by PCR and has shown excellent sensitivity and specificity [65,66].

4. Treatment and Prevention

Due to the different amoebiasis forms and severity of the symptoms, the treatment varies, and diverse drugs can be used. Regarding treatment, amoebiasis is usually treated by amebicides, which are prescribed depending on the severity of the infection [67].

Antiamoebics are the most recommended drugs; these are classified as (1) luminal amebicides which included paromomycin, diloxanide furoate, iodoquinol, and nitazoxanide; they act only on the intestinal lumen and are used to treat amoebic, not dysenteric, colitis, and (2) tissue amebicides such as chloroquine, emetine, tinidazole, and metronidazole, which are the elected treatment in patients with symptomatic intestinal amoebiasis, asymptomatic cyst carriers, and amoebic liver abscesses, due to their rapid intestinal absorption acting at the systemic level [15,68]. Metronidazole is the major drug used in invasive amoebiasis cases; although the toxicity and side effects of this compound are important, the most frequent side effects are nausea, vomiting, headaches, a metallic or bitter taste in the mouth, and some more serious effects such as anorexia, ataxia, and skin rashes/itching. Nevertheless, metronidazole is widely used in the treatment of invasive amoebiasis due to its efficiency and its low cost [69–71].

Likewise, the development of resistance has been reported in some anaerobic pathogens [72,73], but in our experience, there is no clear evidence of resistant strains of *E. histolytica* in patients with amoebic liver abscesses. However, it is important that after systemic treatment of invasive amoebiasis such as amoebic liver abscesses to prescribe a luminal antiamoebic treatment with the purpose of avoiding *E. histolytica*/*E. dispar* cysts luminal infection [15].

On the other hand, the primary measures for the prevention and control of amoebiasis infections remain hygiene habits, use of latrines, hand washing, improvement of water purification systems and adequate hygiene practices to prepare food, and the avoidance of fecal–oral exposure, which could all decrease disease incidence [74,75]. It seems easy to avoid amoebiasis infection by adopting these measures; however, these measures are unattainable when we talk about the fact that this infection is endemic in underdeveloped countries and in poor areas of these countries where the basic infrastructure is not available. Therefore, it is necessary that great efforts, on the part of governments and health systems, be directed to generate the conditions and infrastructure necessary to prevent and reduce the incidence of amoebiasis in developing countries.

Unfortunately, despite the efforts of the scientific community in the field, it has not been possible to obtain a vaccine against this disease so far. However, many recent vaccine development studies appear to be promising [20].

5. Pathogenesis

Pathogenesis of *E. histolytica* is multifactorial, since the virulent molecules of the parasite, as well as the host's immune response, play an important role in the pathogenesis of the disease, causing damage to tissues that facilitate entry to systemic sites [76]. The destructive mechanisms of *E. histolytica* include contact with target cells, cytolysis, and phagocytosis, and the degradation of ingested cells. After contact with trophozoites in the epithelium, the cells increase the paracellular permeability produced by the opening of the TJ junctions and the distortion of the microvilli [77]. Superficial blisters and tiny focal discontinuities appear in the host plasma membrane, displacing and separating neighboring cells and compromising membrane integrity [78].

Entamoeba histolytica usually lodges in the intestine and in about 90 percent of cases, it generates an asymptomatic infection; however, it is not clear why a minority of individuals infected with *E. histolytica* activate a pathogenic phenotype [79]. Throughout many years of study, the participation of pathogenic molecules involved in various causal events due to *E. histolytica* infection has been demonstrated. The most studied physiological events are: (1) the colonization of the intestinal mucus layer, (2) the adherence of trophozoites to the host intestinal epithelium, (3) the interaction with the intracellular junctions of the epithelium, (4) cytolysis, (5) phagocytosis and trogocytosis, (6) the activation of the host's immune response, and (7) the contributions of the host gut microbiota [4,78,80,81]

The genome sequence of *E. histolytica* strain HM1: IMSS was published and analyzed in 2005 [82–84]. The genome assembly contained 20, 800, and 560 bp of DNA in 1496 scaffolds. The genome had a high AT content (approximately 75%). Likewise, approximately half of the assembled sequence was predicted to be coding, with 8333 annotated genes [85]. Since it was sequenced in 2005, the genome of *E. histolytica* has been extensively analyzed due to its implications for human health [86].

These analyses of the *E. histolytica* genome have revealed a variety of metabolic adaptations, which include the reduction in or elimination of most mitochondrial metabolic pathways and the use of oxidative stress enzymes generally associated with anaerobic prokaryotes. Based on the knowledge of the *E. histolytica* genome there is evidence that suggests that cellular events such as meiotic recombination and the possible existence of a sexual reproduction mechanism have occurred in its chromosomes, which could allow the transfer of genetic information between different strains of *E. histolytica*, generating amoebas with a greater capacity to invade tissues and greater resistance to drugs, allowing them to survive in the host and even increase their capacity to survive outside it [87,88].

Another characteristic that has been observed in the *E. histolytica* genome is the so-called gain of genetic information, which includes the significant number of genes (at least 68) that seem to have been gained by horizontal transfer of bacterial genes from intestinal flora, some of them from the same donor group: alpha-1,2-mannosidase (EHI 009520), mannose-1-phosphate guanylyltransferase (EHI 052810), and fructokinase (EHI 054510) from Bacteroidetes; nicotinatephosphoribosyltransferase (EHI 023260) and hypothetical protein (EHI 072640) from Bacteroidales; and Fe-S cluster assembly protein NifU (EHI 049620) and metallo-beta-lactamase family protein (EHI 068560) from proteobacteria [82,84,89]. These genes tend to be involved in metabolic processes characteristic of the anaerobic lifestyle of the organisms [82]. Likewise, the analysis of the *E. histolytica* genome has also revealed that different genes encode proteins belonging to the A1G1 family, whose function is related to drug resistance in bacteria [90]. It has been observed that the expression of this group of proteins is strongly diminished in *E. dispar*; the previous study suggests that these molecules could participate directly in the virulence of the pathogenic amoebae [91]. Some genes that are important in the virulence of *E. histolytica* are genes that encode BspA-like proteins (BspA-like) that participate in cell aggregation by binding to extracellular matrix proteins and in adhesion to target cells; these functions are important for cellular damage and can also stimulate the production of proinflammatory cytokines through the stimulation of Toll-like receptors, which participate in the host innate immune response responsible for limiting tissue damage [92]. These cytokines are believed to help generate an inflammatory microenvironment that could favor the destruction of cells by amoebae [91]. The analysis of the *E. histolytica* genome has also made it possible to identify redundancy in genes that have been widely studied for their participation in different cellular processes related to the pathogenesis of *E. histolytica*. For example, the genes encoding the lectin Gal/GalNAc is important for the adhesion of trophozoites to epithelial cells; the genes that encode for proteins called amoebopores and some of the cysteine proteinases are involved in the cytolysis of epithelial cells [93]. Some of the molecules that have recently been identified, and the mechanisms involved in epigenetic modifications, such as DNA methylation or histone modification in *E. histolytica*, act directly in the conformation of chromatin and the regulation of gene expression [94]. These epigenetic regulatory mechanisms could be important to activate or prevent the expression of genes involved in virulence and cyst formation.

Another recent discovery in the study of the *E. histolytica* genome has been the presence of elements called transposons and retrotransposons, which constitute a significant portion of the genome of *E. histolytica* [95]. The insertion of transposons and retrotransposons results in changes in the amount of DNA, as well as changes and mutation genes. Although until now there is not much information about the function of this type of molecule, it is thought that these elements of amoebic DNA could be participating in the generation of pathogenic variants of *E. histolytica*, which are obtained from clinical isolates and cultured in the presence of pathogenic bacteria, in which there is an increase in the virulence of *E. histolytica*.

Another molecule that participates in *E. histolytica* infection is the cyclooxygenase from *E. histolytica* (EhCox). EhCox, derived from prostaglandin E2, stimulates the chemokine IL-8 from mucosal epithelial cells that recruit neutrophils to the site of infection, increasing the tissue damage. This phenomenon was verified by making gene silencing of EhCox (EhCoxgs) which increases the expression and production of endogenous cysteine protease (CP) and the virulence of trophozoites without altering CP gene transcripts. Shahi et al. propose a negative feedback mechanism in *E. histolytica* to limit proteolytic activity during colonization that can inadvertently trigger inflammation in the gut-enhanced erythrophagocytosis, cytopathic effects on colonic epithelial cells, and elicited proinflammatory cytokines in mice colonic loops, as well as increased inflammation associated with high levels of myeloperoxidase activity [76].

The sphingomyelinase enzymes encoded by *E. histolytica* are other molecules participating in the pathogenesis of three of the neutral sphingomyelinase enzymes encoded in the *E. histolytica* genome. In transfected trophozoites, an overexpression of EhnSM1 and EhnSM2 caused an increase in cytopathic activity and EhnSM3 induced an increase in hemolytic and cytotoxic activities. Another of the most important virulence molecules in *E. histolytica* is the amoebopore, which is a pore-forming protein. An increase in gene expression of amoebapores A, B, and C has been seen in virulent *E. histolytica* strains in comparison with less virulent *Entamoeba* species. Urquieta-Ramírez et al. [96] mention that in a *E. histolytica* strain that overexpresses neutral sphingomyelinases, the gene expression levels for cysteine proteinase 5, adhesin 112, and heavy and light Gal/GalNAc lectin subunits are not affected. They propose that the increase in the cytotoxic and lytic effect of EhnSM3 overexpressing strain can be related to the sum of the effect of EhnSM3 plus amoebapores in a process cell contact-dependent manner or as a mediator by inducing the gene expression of amoebapores enabling the link between EhnSM3 with the virulence phenotype of *E. histolytica*.

Efforts have also been made to elucidate the pathogenicity of *E. histolytica* from genomic studies. Wilson et al. [97] showed differences in the expansion and contraction of families of proteins associated with host or bacterial interactions and highlighted the importance to parasitic *Entamoeba* species of surface-bound proteins involved in adhesion to extracellular membranes, such as the Gal/GalNAc lectin and members of the BspA and Ariel1 families. On the other hand, Nakada-Tsukui et al. [98] find that the AIG1 family protein plays a pivotal role in *E. histolytica* virulence via the regulation of host cell adhesion.

The PTP superfamily is classified as classical tyrosine-specific phosphatases, dual-specificity phosphatases, cdc25 phosphatases, and low-molecular-weight protein tyrosine phosphatase (LMW-PTP) [99]. The genome of *E. histolytica* contains two genes codifying for LMW-PTPs, EhLMW-PTP1, and EhLMW-PTP2, which are identical to the amino acid sequence except for a single conservative residue change at position A85 V. Putative substrates have been identified using an EhLMW-PTP1 mutant at the catalytic site, suggesting it plays an essential role in parasite virulence [100]. Torres-Cifuentes et al. [101] analyzed the potential role of EhLMW-PTP2 in the virulence of *E. histolytica* and found a novel and unexpected effect on adhesion, phagocytosis, and cytopathic effect mediated by regulating the LMW-PTP2 mRNA and protein expression in transfected amoebas. They suggested that EhLMW-PTPs are essential enzymes that regulate cellular pathways activated in the parasite during infection.

On the other hand, it is known that iron is a determinant in the survival of *E. histolytica* and can modulate the expression of virulence factors. Some of the most representative events of the role of iron in the survival of *E. histolytica* are specified by Gastelum-Martínez et al. The most highly expressed virulence factors include EhCP1, EhCP2, EhCP3, and EhCP5. EhCP1-6 is overexpressed in limited iron conditions in culture. Negative regulation by iron was shown for Ehmmbp26 and Ehmmbp45, which obtain iron from the haem group [102]. Aldose reductase participates in siderophore biosynthesis and/or ROS formation. Superoxide dismutase (EhSOD) is responsible for the detoxification of superoxide radicals. Acetaldehyde/alcohol dehydrogenase-2 (EhADH2) is responsible for the internalization of human holo-transferrin. NADPH-dependent oxidoreductase (EhNO2) is involved in redox homeostasis through the reduction of L-cysteine and iron. Actin and genes related to the reorganization of the parasite cytoskeleton such as actobindin and cofilin are downregulated by iron [103]. The iron concentration could affect the cytoadherence capacity of the trophozoite. Acyl-CoA synthetase is involved in the catabolism of fatty acids. Cysteine synthase CS1 and CS2 and cysteine desulphurase (NifS) are involved in cysteine metabolism and possibly in the metabolism of sugar-containing amino acids, and both are upregulated by iron. S-adenosylmethionine synthetase (SAM) participates in the biosynthesis of SAM and many biosynthetic pathways including methionine-cysteine conversion and haem biosynthesis. SAM is downregulated by iron. The presence of cis-elements in mRNA (iron-responsive elements (IREs)) was demonstrated in the actin and Ehhmb26 genes, suggesting the presence of a post-transcriptional IRE/iron-responsive protein [IRP] iron regulatory system in this parasite.

During chronic inflammation, the vagus nerve limits the immune response through the anti-inflammatory reflex, which includes acetylcholine (ACh) as one of the predominant neurotransmitters at the infection site. Consequently, the response of *E. histolytica* trophozoites to ACh could be implicated in the establishment of invasive disease. Medina-Rosales et al. [104] evaluated the effect of ACh on *E. histolytica* virulence. They demonstrated that *E. histolytica* trophozoites bind ACh on their membrane and show a clear increase in the expression of virulence factors that were upregulated upon stimulation with the neurotransmitter. ACh treatment increased the expression of L220, Gal/GalNAc lectin heavy subunit (170 kDa), amoebapore C, cysteine proteinase 2 (ehcp-a2), and cysteine proteinase 5 (ehcp-a5). Moreover, erythrophagocytosis, cytotoxicity, and actin cytoskeleton remodeling were augmented after ACh treatment.

Lipids are essential in the cellular processes and pathogenesis of this organism *Entamoeba histolytica* has 27 putative enzymes involved in the lipid metabolic pathways and the trophozoites have mechanisms to synthesize or remodel phospholipids. They acquire cholesterol and fatty acids from the environment; thus, parasites need lipid scavenging from the host to regulate their cellular processes for survival or they may provoke disequilibrium in lipid synthesis and metabolism in the organism. However, lipid metabolism during host–pathogen interaction research is still an open field to be explored and exploited for the development of new methodologies for diagnosis and treatment [17].

6. Intestinal Ecosystem and *Entamoeba* Interactions

This is a very recent field of research and there are highly relevant findings on the interaction of *E. histolytica* in the colon with multiple components that modulate the invasion and destruction of tissues in the outer mucus layer [105]. In the colon, *E. histolytica* interacts with a diverse microbiota and takes advantage of its presence since it uses its decomposition products (such as glycans) that can serve as a source of nutrients, and at the same time, *E. histolytica* can feed on the colonic microbiota [105,106]. Recently, a metagenomics study on the in vitro association between *E. histolytica* and enteric bacteria showed that *E. histolytica* preferentially phagocytoses *Lactobacillus ruminus*, *Faecalibacterium prausnitzii*, *Bifidobacterium longum*, and *B. ruminantium*. These findings suggest that *E. histolytica* prefers to phagocytose commensal bacteria that are part of the healthy microbiota [106].

In addition to the predator–prey relationship, there is recent evidence that when the infection caused by *E. histolytica* is symptomatic, the intestinal microbiota could play an important role in the development of the disease [9,107]. Reyna-Fabian et al. examined samples aspirated from abscess material obtained from patients who were clinically diagnosed with amoebic liver abscesses (ALAs) or pyogenic liver abscesses (PLAs). They suggested that most of the ALAs or PLAs are mixed abscesses (bacteria and protozoa) and that the disease outcomes could be related to the bacterial/protozoan density ratio [35]. This is in line with Ximénez et al., 2010, who hypothesized that the size of intestinal amebic ulcers and the local inflammatory response permits the exit of *E. histolytica* trophozoites and intestinal bacteria, some of which enter the phagocytic vacuoles of trophozoites [27].

It has been determined in experimental models, that both the pathogenicity and virulence of *E. histolytica* are affected by the presence or absence of enteric bacteria and this could help explain why only about 10% of individuals infected with *E. histolytica* develop intestinal amoebiasis [105,108–110]. Evaluating the pathogenicity of *E. histolytica* with co-infection with other parasites has been reported in in vitro studies to show a twofold increase in CaCo monolayer destruction when *E. histolytica* interacted with enteropathogenic *Escherichia coli* (EPEC), but not with *E. coli* DH5α, for 2.5 h. This was associated with increased *E. histolytica* proteolytic activity as revealed by zymogram analysis and degradation of the *E. histolytica* CP-A1/5 (EhCP-A1/5) peptide substrate Z-Arg-Arg-pNC and EhCP4 substrate Z-Val-Val-Arg-AMC. Additionally, *E. histolytica*–EPEC interaction increased EhCP-A1, -A2, -A4, -A5, Hgl, Apa, and Cox-1 mRNA expression. In animal models of disease, the *E. histolytica*–EPEC interaction enhanced cellular inflammatory reaction, granuloma formation, and necrosis in ALA in hamsters and increased secretory and proinflammatory cytokine responses in closed colonic loops in mice [111].

Furthermore, Reyna-Fabian (2010) [35] showed that *E. dispar* in association with the gut microbiota could be potentially responsible for intestinal or liver tissue damage, like that observed in *E. histolytica*. More recently, Vilela da Silva et al. (2021) reported new evidence about some *E. dispar* strains in that the cysteine proteinase 5 expression in *E. dispar* MCR, VEJ, and ADO strains, isolated in Brazil, show that *S. typhimurium* and *E. dispar* co-infection worsens amoebic colitis, possibly by increasing the expression of amoebic virulence factors [112]. However, great effort is still required to elucidate the factors that drove these interactions relevant to the development of the disease caused by *E. dispar*.

On the other hand, important alterations in the intestinal microbiota composition have been observed due to infection by E. histolytica [9]. A dysbiotic state has been observed in linked E. histolytica-positive patients in India, characterized by a decrease in *Bacteroides*, *Clostridium coccoides*, *C. leptum*, *Lactobacillus*, *Campylobacter*, and *Eubacterium* with a corresponding increase in *Bifidobacterium* species [105,113]. Another recent study in Cameroon also linked the presence of *E. histolytica* with an increase in *Clostridial* and *Ruminococcaceae* and a corresponding decrease in *Prevotella copri* and *Fusobacteria* [114]. This study almost showed that in *E. histolytica*-infected individuals an augmented number of bacterial species (alpha diversity) and a decrease in interindividual variation (beta diversity) are evident. Also, it was recently demonstrated that a dysbiotic state renders the host hyper-responsive for increased proinflammatory cytokine and chemokine production with hypersecretory responses toward *E. histolytica* [105]. The authors mentioned that these findings are of great relevance, as individuals with dysbiosis (either by disease, antibiotics, or due to a poor diet) that are infected with *E. histolytica* are at a high risk of developing severe intestinal amoebiasis associated with acute inflammation, compared to individuals with a healthy microbiota. The fact that *E. histolytica* induces the production of antimicrobial peptides but is resistant to their cytopathic effect could explain the alteration in microbial composition observed in *E. histolytica* infections [115].

However, until now, it has been unknown whether the dysbiosis observed in amoebiasis is the cause or the effect of infection by *E. histolytica*, mainly in rural or suburban areas where the indiscriminate use of antibiotics can secondarily induce intestinal dysbiosis; therefore, interactions between *E. histolytica* and the microbiota are an interesting field of research.

7. Discussion and Future Perspectives

Although *E. histolytica* was discovered over 100 years ago, amoebiasis remains a major public health challenge around the world, especially in developing countries where a large segment of the population does not have easy access to standard sanitary conditions and, therefore, does not have good hygienic conditions, making strategies that minimize its transmission ineffective [4]. The only reservoir of *E. histolytica* is humans and the infection occurs through food, water, or hands contaminated with fecal matter that contains cysts. Therefore, it is necessary to develop government prevention and control programs of transmissibility, morbidity, and mortality caused by *E. histolytica* in endemic countries, which is a real challenge. Likewise, as mentioned in this review, person-to-person transmission also occurs through oral–genital and oral–anal contact, especially among the homosexual population, so it is necessary to implement informational strategies to avoid the transmissibility of the different forms of amoebiasis in this population.

Given the different ways amoebiasis can occur, it is important to have robust tests to help make an accurate diagnosis and overcome the limitations of the current diagnosis, especially in endemic areas, where individuals are constantly exposed to *E. histolytica*.

Advances in molecular diagnostic methodologies have been of great importance to recognize and differentiate *E. histolytica* from nonpathogenic *Entamoeba* species, and this differentiation has had a great impact on the epidemiology and clinical management of patients. However, although molecular detection is highly sensitive and specific and has the potential to be developed into field-applicable tests, the cost is still a barrier to their extensive usage as a routine laboratory test method in most endemic areas, where a correct diagnosis of amoebic infection is important to avoid morbidity, death, and disease transmission. As well, new strategies must include improving the potency of existing amebicides, discovering the new usage of approved drugs, drug rediscovery, and drug targeting of essential *E. histolytica* components to use in deworming campaigns, especially in endemic countries. Further research areas that need to be explored include the discovery of biomarkers that can predict individuals prone to symptomatic infection, distinguish between the different stages of infection, and monitor treatment success. In the same way, efforts should also be directed at studies that focus on the interactions of *E. histolytica* with microorganisms in the intestinal environment. This knowledge would have a positive impact on the clinical and laboratory diagnosis of diarrheal syndrome, its treatment, and, subsequently, the implementation of more reliable control schemes. Additionally, more research will be needed to understand the relationship between the commensal microbiota and the immune response in this parasite infection. Likewise, the advance in molecular techniques could be an important tool in the study of the geographic distribution of species of *Entamoeba* and could determine the morbidity rates of different forms of amoebic infection in different geographic areas. Although the analysis of the *Entamoeba* genomes has shed light on the study of these parasites, more research is still needed to understand, in more detail, the evolutive history of these microorganisms, in particular *E. histolytica*. In conclusion, understanding the epidemiology, pathology, and molecular biology of this parasite will not only lead to the improvement of diagnostic and treatment options but also, ultimately, to the development of safe and efficacious vaccines for the control of the disease. Finally, and no less important, the improvement of water purification systems and hygiene practices can greatly decrease disease incidence; however, it requires considerable time, changes, and monetary investments.

Author Contributions: P.M., A.S.-V. and C.X. conceptualized the review content and wrote the first draft of the manuscript. L.R.-V., E.G., H.P.-J., T.P.-B. and M.R. edited the manuscript. E.G.H., M.d.l.A.P. and M.E.Z. participated in specific paragraphs and re-reading the text. All authors contributed to the revision of the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: The present work was supported by Grants from PAPIIT IN217821 and IN200420 (DGAPA), Universidad Nacional Autónoma de México (UNAM). CONACyT-Salud 272601 S008-Fonsec SSA/IMSS/ISSSTE.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors highly appreciate Miriam Nieves-Ramírez for her assistance in the pathogenesis section of the manuscript. H.P.-J. received a research grant from the Postdoctoral Stays in Mexico, program-CONACyT, for his postdoctoral stay (CVU 408467).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zhang, Y.; Niu, Q.; Fan, W.; Huang, F.; He, H. Oral Microbiota and Gastrointestinal Cancer. *OTT* **2019**, *12*, 4721–4728. [CrossRef]
2. WHO/PAHO/UNESCO Report. A Consultation with Experts on Amoebiasis. Mexico City, Mexico, 28–29 January 1997. *Epidemiol. Bull.* **1997**, *18*, 13–14.
3. Ximénez, C.; Morán, P.; Rojas, L.; Valadez, A.; Gómez, A. Reassessment of the Epidemiology of Amebiasis: State of the Art. *Infect. Genet. Evol.* **2009**, *9*, 1023–1032. [CrossRef]

4. Carrero, J.C.; Reyes-López, M.; Serrano-Luna, J.; Shibayama, M.; Unzueta, J.; León-Sicairos, N.; De La Garza, M. Intestinal Amoebiasis: 160 Years of Its First Detection and Still Remains as a Health Problem in Developing Countries. *Int. J. Med. Microbiol.* **2020**, *310*, 151358. [CrossRef]
5. Nath, J.; Ghosh, S.K.; Singha, B.; Paul, J. Molecular Epidemiology of Amoebiasis: A Cross-Sectional Study among North East Indian Population. *PLoS Negl. Trop. Dis.* **2015**, *9*, e0004225. [CrossRef]
6. Cui, Z.; Li, J.; Chen, Y.; Zhang, L. Molecular Epidemiology, Evolution, and Phylogeny of *Entamoeba* spp. *Infect. Genet. Evol.* **2019**, *75*, 104018. [CrossRef]
7. Tharmaratnam, T.; Kumanan, T.; Iskandar, M.A.; D'Urzo, K.; Gopee-Ramanan, P.; Loganathan, M.; Tabobondung, T.; Tabobondung, T.A.; Sivagurunathan, S.; Patel, M.; et al. *Entamoeba histolytica* and Amoebic Liver Abscess in Northern Sri Lanka: A Public Health Problem. *Trop. Med. Health* **2020**, *48*, 2. [CrossRef]
8. Shirley, D.-A.; Moonah, S. Fulminant Amebic Colitis after Corticosteroid Therapy: A Systematic Review. *PLoS Negl. Trop. Dis.* **2016**, *10*, e0004879. [CrossRef]
9. Ankri, S. *Entamoeba histolytica*—Gut Microbiota Interaction: More Than Meets the Eye. *Microorganisms* **2021**, *9*, 581. [CrossRef]
10. Nagaraja, S.; Ankri, S. Target Identification and Intervention Strategies against Amebiasis. *Drug Resist. Updates* **2019**, *44*, 1–14. [CrossRef]
11. Ximénez, C.; Morán, P.; Rojas, L.; Valadez, A.; Gómez, A.; Ramiro, M.; Cerritos, R.; González, E.; Hernández, E.; Oswaldo, P. Novelties on Amoebiasis: A Neglected Tropical Disease. *J. Glob. Infect. Dis.* **2011**, *3*, 166. [CrossRef]
12. Mi-ichi, F.; Ikeda, K.; Tsugawa, H.; Deloer, S.; Yoshida, H.; Arita, M. Stage-Specific De Novo Synthesis of Very-Long-Chain Dihydroceramides Confers Dormancy to *Entamoeba* Parasites. *mSphere* **2021**, *6*, e00174-21. [CrossRef] [PubMed]
13. Vaisusuk, K.; Saijuntha, W. Intestinal Protozoa: Their Role as Human Pathogens and Zoonoses. In *Biodiversity of Southeast Asian Parasites and Vectors Causing Human Disease*; Petney, T.N., Saijuntha, W., Mehlhorn, H., Eds.; Parasitology Research Monographs; Springer Nature Switzerland AG: Cham, Switzerland, 2021; pp. 35–61. ISBN 978-3-030-71161-0.
14. Nakada-Tsukui, K.; Nozaki, T. Immune Response of Amebiasis and Immune Evasion by *Entamoeba histolytica*. *Front. Immunol.* **2016**, *7*, 175. [CrossRef]
15. Shirley, D.-A.T.; Farr, L.; Watanabe, K.; Moonah, S. A Review of the Global Burden, New Diagnostics, and Current Therapeutics for Amebiasis. *Open Forum Infect. Dis.* **2018**, *5*, ofy161. [CrossRef]
16. Ali, I.K.M.; Clark, C.G.; Petri, W.A. Molecular Epidemiology of Amebiasis. *Infect. Genet. Evol.* **2008**, *8*, 698–707. [CrossRef]
17. Castellanos-Castro, S.; Bolaños, J.; Orozco, E. Lipids in *Entamoeba histolytica*: Host-Dependence and Virulence Factors. *Front. Cell. Infect. Microbiol.* **2020**, *10*, 75. [CrossRef]
18. Shirley, D.-A.; Hung, C.-C.; Moonah, S. Entamoeba histolytica (Amebiasis). In *Hunter's Tropical Medicine and Emerging Infectious Disease*; Ryan, E.T., Hunter, G.W., Hill, D.R., Solomon, T., Endy, T.P., Aronson, N., Ryan, E.T., Eds.; Elsevier: Amsterdam, The Netherlands, 2019; pp. 699–706. ISBN 978-0-323-55512-8.
19. Roediger, R.; Lisker-Melman, M. Pyogenic and Amebic Infections of the Liver. *Gastroenterol. Clin. N. Am.* **2020**, *49*, 361–377. [CrossRef]
20. Kantor, M.; Abrantes, A.; Estevez, A.; Schiller, A.; Torrent, J.; Gascon, J.; Hernandez, R.; Ochner, C. *Entamoeba histolytica*: Updates in Clinical Manifestation, Pathogenesis, and Vaccine Development. *Can. J. Gastroenterol. Hepatol.* **2018**, *2018*, 4601420. [CrossRef]
21. Victoria-Hernández, J.A.; Ventura-Saucedo, A.; López-Morones, A.; Martínez-Hernández, S.L.; Medina-Rosales, M.N.; Muñoz-Ortega, M.; Ávila-Blanco, M.E.; Cervantes-García, D.; Barba-Gallardo, L.F.; Ventura-Juárez, J. Case Report: Multiple and Atypical Amoebic Cerebral Abscesses Resistant to Treatment. *BMC Infect. Dis.* **2020**, *20*, 669. [CrossRef]
22. El-Radhi, A.S. Fever in Common Infectious Diseases. In *Clinical Manual of Fever in Children*; El-Radhi, A.S., Ed.; Springer International Publishing: Cham, Switzerland, 2018; pp. 85–140. ISBN 978-3-319-92335-2.
23. Martínez-Palomo, A.; Espinosa-Cantellano, M.; Tsutsumi, G.R.; Tsutsumi, V. Amebic Liver Abscess. In *Liver Pathophysiology: Therapies and Antioxidants*; Muriel, P., Ed.; Academic Press: London, UK, 2017; pp. 181–186. ISBN 978-0-12-804321-9.
24. Tanaka, E.; Tashiro, Y.; Kotake, A.; Takeyama, N.; Umemoto, T.; Nagahama, M.; Hashimoto, T. Spectrum of CT Findings in Amebic Colitis. *Jpn. J. Radiol.* **2021**, *39*, 558–563. [CrossRef] [PubMed]
25. Rojas, L.; Morán, P.; Valadez, A.; Gómez, A.; González, E.; Hernández, E.; Partida, O.; Nieves, M.; Gudiño, M.; Magaña, U.; et al. *Entamoeba histolytica* and *Entamoeba dispar* Infection in Mexican School Children: Genotyping and Phylogenetic Relationship. *BMC Infect. Dis.* **2016**, *16*, 485. [CrossRef]
26. Ximenez, C.; Partida, O.; Nieves, M.; Hernandez, E.; Moran, P.; Valadez, A.; Gonzalez, E.; Cerritos, R.; Rojas, L. Immune Response in Human Amebiasis: A Protective Response? In *Amebiasis: Biology and Pathogenesis of Entamoeba*; Springer: Berlin/Heidelberg, Germany, 2015.
27. Ximénez, C.; Cerritos, R.; Rojas, L.; Dolabella, S.; Morán, P.; Shibayama, M.; González, E.; Valadez, A.; Hernández, E.; Valenzuela, O.; et al. Human Amebiasis: Breaking the Paradigm? *Int. J. Environ. Res. Public Health* **2010**, *7*, 1105–1120. [CrossRef] [PubMed]
28. Guarner, V. Treatment and Prevention of Amebiasis. In *Amebiasis: Infection and Disease by Entamoeba Histolytica*; Kretschmer, R.R., Ed.; CRC Press: Boca Raton, FL, USA, 2019; pp. 221–238. ISBN 978-0-429-28253-9.
29. Proctor, E.M. Laboratory Diagnosis of Amebiasis. *Clin. Lab. Med.* **1991**, *11*, 829–859. [CrossRef]
30. Tanyuksel, M.; Petri, W.A. Laboratory Diagnosis of Amebiasis. *Clin. Microbiol. Rev.* **2003**, *16*, 713–729. [CrossRef] [PubMed]
31. Saidin, S.; Othman, N.; Noordin, R. Update on Laboratory Diagnosis of Amoebiasis. *Eur. J. Clin. Microbiol. Infect. Dis.* **2019**, *38*, 15–38. [CrossRef]

32. Juwita, P.M.; Yudhawati, R. Secondary Pleuropulmonary Amoebiasis Due to Liver Abscess Rupture: A Complication Case Report in Low Resource Setting. *Int. J. Surg. Case Rep.* **2021**, *85*, 106231. [CrossRef]
33. Pritt, B.S.; Clark, C.G. Amebiasis. *Mayo Clin. Proc.* **2008**, *83*, 1154–1160. [CrossRef]
34. Dolabella, S.S.; Serrano-Luna, J.; Navarro-García, F.; Cerritos, R.; Ximénez, C.; Galván-Moroyoqui, J.M.; Silva, E.F.; Tsutsumi, V.; Shibayama, M. Amoebic Liver Abscess Production by *Entamoeba dispar*. *Ann. Hepatol.* **2012**, *11*, 107–117. [CrossRef]
35. Reyna-Fabián, M.E.; Zermeño, V.; Ximénez, C.; Flores, J.; Romero, M.F.; Diaz, D.; Argueta, J.; Moran, P.; Valadez, A.; Cerritos, R. Analysis of the Bacterial Diversity in Liver Abscess: Differences Between Pyogenic and Amebic Abscesses. *Am. J. Trop. Med. Hyg.* **2016**, *94*, 147–155. [CrossRef]
36. Sayek, I.; Onat, D. Pyogenic and Amebic Liver Abscess. In *Surgical treatment: Evidence-Based and Problem-Oriented*; Holzheimer, R., Mannick, J.A., Eds.; Zuckschwerdt: München, Germeny, 2001; ISBN 978-3-88603-714-8.
37. Rigothier, M.C.; Khun, H.; Tavares, P.; Carmona, A.; Huerre, M.; Guillén, N. Fate of *Entamoeba histolytica* during Establishment of Amoebic Liver Abscess Analyzed by Quantitative Radioimaging and Histology. *Infect. Immun.* **2002**, *2002 70*, 3208–3215. [CrossRef]
38. Pandak, N.; Golchinheydari, S.; Mahdi, A.S.; Al Majrafi, A.; Deenadayalan, S.S.; Khamis, F. Amebic Liver Abscess: A Disease Native to Oman? *Sultan. Qaboos. Univ. Med. J.* **2021**, *22*, 253. [CrossRef]
39. Moran, P.; Gómez, A.; Valadez, A.; Ramos, F.; García, G.; Limón, A.; Valenzuela, O.; Ramiro, M.; Melendro, E.I.; Ximenez, C. Amebic and Pyogenic Liver Abscess: Importance of Differential Diagnosis in Endemic Areas of Amebiasis. *Trop. Med. Int. Health* **2007**, *12*, 153–154. [CrossRef]
40. Sánchez-Guillén, M.D.C.; Pérez-Fuentes, R.; Salgado-Rosas, H.; Ruiz-Argüelles, A.; Ackers, J.; Shire, A.; Talamas-Rohana, P. Differentiation of *Entamoeba histolytica*/*Entamoeba dispar* by PCR and Their Correlation with Humoral and Cellular Immunity in Individuals with Clinical Variants of Amoebiasis. *Am. J. Trop. Med. Hyg.* **2002**, *66*, 731–737. [CrossRef]
41. López Revilla, R.; Navarro-García, F.; Valadez-Sánchez, M.; López Vidal, Y.; Calva Mercado, J. Dot-Enzyme-Linked Immunosorbent Assay (Dot-ELISA) of Anti-*Entamoeba histolytica* Antibodies in Human Serum and Colostrum. *Arch. Invest. Med. Mex* **1991**, *22*, 249–253.
42. Schouten, W.J.; Oostburg, B.F.J.; Noordpool, H.; Makbin, M.; Bos, H.J. A Seroepidemiological Study of Amebiasis in Surinam by the Enzyme-Linked Immunosorbent Assay (Elisa)*. *Am. J. Trop. Med. Hyg.* **1980**, *29*, 358–363. [CrossRef]
43. Salvatierra, B.; Tapia-Conyer, R.; Sepulveda-Amor, J.; Gutierrez, G.; Ortiz-Ortiz, L.; Caballero-Salcedo, A.; Viveros-Rogel, M. Seroepidemiology of Amebiasis in Mexico. *Am. J. Trop. Med. Hyg.* **1994**, *50*, 412–419. [CrossRef]
44. Haque, R.; Neville, L.M.; Hahn, P.; Petri, W.A. Rapid Diagnosis of *Entamoeba* Infection by Using *Entamoeba* and *Entamoeba histolytica* Stool Antigen Detection Kits. *J. Clin. Microbiol.* **1995**, *33*, 2558–2561. [CrossRef]
45. Haque, R.; Kress, K.; Wood, S.; Jackson, T.F.H.G.; Lyerly, D.; Wilkins, T.; Petri, W.A. Diagnosis of Pathogenic *Entamoeba histolytica* Infection Using a Stool ELISA Based on Monoclonal Antibodies to the Galactose-Specific Adhesin. *J. Infect. Dis.* **1993**, *167*, 247–249. [CrossRef] [PubMed]
46. Haque, R.; Petri, W.A. Diagnosis of Amebiasis in Bangladesh. *Arch. Med. Res.* **2006**, *37*, 272–275. [CrossRef]
47. Spadafora, L.J.; Kearney, M.R.; Siddique, A.; Ali, I.K.; Gilchrist, C.A.; Arju, T.; Hoffstrom, B.; Nguyen, F.K.; Petri, W.A.; Haque, R.; et al. Species-Specific Immunodetection of an *Entamoeba histolytica* Cyst Wall Protein. *PLoS Negl. Trop. Dis.* **2016**, *10*, e0004697. [CrossRef]
48. Rivera, W.L.; Tachibana, H.; Silva-Tahat, M.R.A.; Uemura, H.; Kanbara, H. Differentiation of *Entamoeba histolytica* and *E. dispar* DNA from Cysts Present in Stool Specimens by Polymerase Chain Reaction: Its Field Application in the Philippines. *Parasitol. Res.* **1996**, *82*, 585–589. [CrossRef]
49. Clark, C.G.; Diamond, L.S. Intraspecific Variation and Phylogenetic Relationships in the Genus *Entamoeba* as Revealed by Riboprinting. *J. Eukaryot. Microbiol.* **1997**, *44*, 142–154. [CrossRef]
50. Fotedar, R.; Stark, D.; Beebe, N.; Marriott, D.; Ellis, J.; Harkness, J. PCR Detection of *Entamoeba histolytica*, *Entamoeba dispar* and *Entamoeba moshkovskii* in Stool Samples from Sydney, Australia. *J. Clin. Microbiol.* **2007**, *45*, 1035–1037. [CrossRef] [PubMed]
51. Troll, H.; Marti, H.; Weiss, N. Simple Differential Detection of *Entamoeba histolytica* and *Entamoeba dispar* in Fresh Stool Specimens by Sodium Acetate- Acetic Acid-Formalin Concentration and PCR. *J. Clin. Microbiol.* **1997**, *35*, 1701–1705. [CrossRef]
52. Ramos, F.; Morán, P.; González, E.; García, G.; Ramiro, M.; Gómez, A.; De León, M.D.C.G.; Melendro, E.I.; Valadez, A.; Ximénez, C. *Entamoeba histolytica* and *Entamoeba dispar*: Prevalence Infection in a Rural Mexican Community. *Exp. Parasitol.* **2005**, *110*, 327–330. [CrossRef] [PubMed]
53. Hamzah, Z.; Petmitr, S.; Mungthin, M.; Leelayoova, S.; Chavalitshewinkoon-Petmitr, P. Differential Detection of *Entamoeba histolytica*, *Entamoeba dispar* and *Entamoeba moshkovskii* by a Single-Round PCR Assay. *J. Clin. Microbiol.* **2006**, *44*, 3196–3200. [CrossRef]
54. Fonte, L.; Montano, I.; Núñez, Y.O.; Maestre, J.L.; Fernández, M.A.; Torres-Núñez, D.; Silva, J.A. Multiplex Polymerase Chain Reaction Amplification and Differentiation of *Entamoeba histolytica* and *Entamoeba dispar* DNA from Stool Samples. *Am. J. Trop. Med. Hyg.* **2001**, *64*, 293–297. [CrossRef]
55. Khairnar, K.; Parija, S.C. A Novel Nested Multiplex Polymerase Chain Reaction (PCR) Assay for Differential Detection of *Entamoeba histolytica*, E. Moshkovskii and E. Dispar DNA in Stool Samples. *BMC Microbiol.* **2007**, *7*, 47. [CrossRef] [PubMed]

56. Mirelman, D.; Nuchamowitz, Y.; Stolarsky, T. Comparison of Use of Enzyme-Linked Immunosorbent Assay-Based Kits and PCR Amplification of RRNA Genes for Simultaneous Detection of *Entamoeba histolytica* and *E. dispar*. *J. Clin. Microbiol.* **1997**, *35*, 2405–2407. [CrossRef] [PubMed]
57. Millan-Velasco, F.; Schoolnick, G.; Wirth, D.; Acuna-Soto, R.; Samuelson, J.; De Girolami, P.; Zarate, L. Application of the Polymerase Chain Reaction to the Epidemiology of Pathogenic and Nonpathogenic *Entamoeba histolytica*. *Am. J. Trop. Med. Hyg.* **1993**, *48*, 58–70. [CrossRef]
58. Evangelopoulos, G.; Spanakos, E.; Pat, A. A Nested, Multiplex, PCR Assay for the Simultaneous Detection and Differentiation of *Entamoeba histolytica* and *Entamoeba dispar* in Faeces. *Ann. Trop. Med. And. Parasitol.* **2000**, *94*, 233–240. [CrossRef]
59. Myjak, P.; Kur, J.; Pietkiewicz, H. Usefulness of New DNA Extraction Procedure for PCR Technique in Species Identification of Entamoeba Isolates. *Wiad. Parazytol.* **1997**, *43*, 163–170.
60. Verweij, J.J.; Blotkamp, J.; Brienen, E.A.T.; Aguirre, A.; Polderman, A.M. Differentiation of *Entamoeba histolytica* and *Entamoeba dispar* Cysts Using Polymerase Chain Reaction on DNA Isolated from Faeces with Spin Columns. *Eur. J. Clin. Microbiol. Infect. Dis.* **2000**, *19*, 358–361. [CrossRef] [PubMed]
61. Verweij, J.J.; Blangé, R.A.; Templeton, K.; Schinkel, J.; Brienen, E.A.T.; Van Rooyen, M.A.A.; Van Lieshout, L.; Polderman, A.M. Simultaneous Detection of *Entamoeba histolytica*, *Giardia lamblia*, and *Cryptosporidium parvum* in Fecal Samples by Using Multiplex Real-Time PCR. *J. Clin. Microbiol.* **2004**, *42*, 1220–1223. [CrossRef] [PubMed]
62. Roy, S.; Kabir, M.; Mondal, D.; Ali, I.K.M.; Petri, W.A.; Haque, R. Real-Time-PCR Assay for Diagnosis of *Entamoeba histolytica* Infection. *J. Clin. Microbiol.* **2005**, *43*, 2168–2172. [CrossRef]
63. Stark, D.; Van Hal, S.; Fotedar, R.; Butcher, A.; Marriott, D.; Ellis, J.; Harkness, J. Comparison of Stool Antigen Detection Kits to PCR for Diagnosis of Amebiasis. *J. Clin. Microbiol.* **2008**, *46*, 1678–1681. [CrossRef]
64. Stark, D.; Fotedar, R.; van Hal, S.; Beebe, N.; Marriott, D.; Ellis, J.T.; Harkness, J. Prevalence of Enteric Protozoa in Human Immunodeficiency Virus (HIV)-Positive and HIV-Negative Men Who Have Sex with Men from Sydney, Australia. *Am. J. Trop. Med. Hyg.* **2007**, *76*, 549–552. [CrossRef] [PubMed]
65. Tachibana, H.; Kobayashi, S.; Okuzawa, E.; Masuda, G. Detection of Pathogenic *Entamoeba histolytica* DNA in Liver Abscess Fluid by Polymerase Chain Reaction. *Int. J. Parasitol.* **1992**, *22*, 1193–1196. [CrossRef]
66. Zaman, S.; Khan, M.A. Direct Ampli®cation of *Entamoeba histolytica* DNA from Amoebic Liver Abscess Pus Using Polymerase Chain Reaction. *Parasitol. Res.* **2000**, *86*, 724–728. [CrossRef]
67. Martínez-Castillo, M.; Pacheco-Yepez, J.; Flores-Huerta, N.; Guzmán-Téllez, P.; Jarillo-Luna, R.A.; Cárdenas-Jaramillo, L.M.; Campos-Rodríguez, R.; Shibayama, M. Flavonoids as a Natural Treatment Against *Entamoeba histolytica*. *Front. Cell. Infect. Microbiol.* **2018**, *8*, 209. [CrossRef]
68. Gonzales, M.L.M.; Dans, L.F.; Sio-Aguilar, J. Antiamoebic Drugs for Treating Amoebic Colitis. *Cochrane Database Syst. Rev.* **2019**, *1*, CD006085. [CrossRef] [PubMed]
69. Haque, R.; Huston, C.D.; Hughes, M.; Houpt, E.; Petri, W.A. Amebiasis. *N. Engl. J. Med.* **2003**, *348*, 1565–1573. [CrossRef] [PubMed]
70. Marie, C.; Petri, W.A. Amoebic Dysentery. *BMJ Clin. Evid.* **2013**, *2013*, 0918. [PubMed]
71. Ansari, M.F.; Siddiqui, S.M.; Agarwal, S.M.; Vikramdeo, K.S.; Mondal, N.; Azam, A. Metronidazole Hydrazone Conjugates: Design, Synthesis, Antiamoebic and Molecular Docking Studies. *Bioorg. Med. Chem. Lett.* **2015**, *25*, 3545–3549. [CrossRef]
72. Adagu, I.S. In Vitro Activity of Nitazoxanide and Related Compounds against Isolates of *Giardia intestinalis*, *Entamoeba histolytica* and *Trichomonas vaginalis*. *J. Antimicrob. Chemother.* **2002**, *49*, 103–111. [CrossRef]
73. Bansal, D.; Sehgal, R.; Chawla, Y.; Mahajan, R.C.; Malla, N. In Vitro Activity of Antiamoebic Drugs against Clinical Isolates of *Entamoeba histolytica* and *Entamoeba dispar*. *Ann. Clin. Microbiol. Antimicrob.* **2004**, *3*, 27. [CrossRef]
74. Farthing, M.J. Treatment Options for the Eradication of Intestinal Protozoa. *Nat. Rev. Gastroenterol. Hepatol.* **2006**, *3*, 436–445. [CrossRef]
75. Souza Passos, A.; Da Silva Santos, R.; Dos Santos Andrade, R.; Ribeiro Santos Junior, E.; Santos Bomfim, E.; Sousa Ribeiro, L.; Andrade Dos Santos, G.; Araujo De Albuquerque, W.; Amor, A.L.M. Enteric Parasites and Socio-Epidemiological Variables in an Academic Community. *Rev. Patol. Trop.* **2021**, *50*, 163–178. [CrossRef]
76. Shahi, P.; Moreau, F.; Chadee, K. Entamoeba histolytica Cyclooxygenase-Like Protein Regulates Cysteine Protease Expression and Virulence. *Front. Cell. Infect. Microbiol.* **2019**, *8*, 447. [CrossRef]
77. Betanzos, A.; Javier-Reyna, R.; García-Rivera, G.; Bañuelos, C.; González-Mariscal, L.; Schnoor, M.; Orozco, E. The EhCPADH112 Complex of *Entamoeba histolytica* Interacts with Tight Junction Proteins Occludin and Claudin-1 to Produce Epithelial Damage. *PLoS ONE* **2013**, *8*, e65100. [CrossRef]
78. Betanzos; Bañuelos; Orozco Host Invasion by Pathogenic Amoebae: Epithelial Disruption by Parasite Proteins. *Genes.* **2019**, *10*, 618. [CrossRef]
79. Begum, S.; Gorman, H.; Chadha, A.; Chadee, K. Role of Inflammasomes in Innate Host Defense against *Entamoeba histolytica*. *J. Leukoc. Biol.* **2020**, *108*, 801–812. [CrossRef]
80. Que, X.; Reed, S.L. Cysteine Proteinases and the Pathogenesis of Amebiasis. *Clin. Microbiol. Rev.* **2000**, *13*, 196–206. [CrossRef] [PubMed]
81. Rawat, A.; Roy, M.; Jyoti, A.; Kaushik, S.; Verma, K.; Srivastava, V.K. Cysteine Proteases: Battling Pathogenic Parasitic Protozoans with Omnipresent Enzymes. *Microbiol. Res.* **2021**, *249*, 126784. [CrossRef]

82. Loftus, B.; Anderson, I.; Davies, R.; Alsmark, U.C.M.; Samuelson, J.; Amedeo, P.; Roncaglia, P.; Berriman, M.; Hirt, R.P.; Mann, B.J.; et al. The Genome of the Protist Parasite *Entamoeba histolytica*. *Nature* **2005**, *433*, 865–868. [CrossRef]
83. Clark, C.G.; Alsmark, U.C.M.; Tazreiter, M.; Saito-Nakano, Y.; Ali, V.; Marion, S.; Weber, C.; Mukherjee, C.; Bruchhaus, I.; Tannich, E.; et al. Structure and Content of the *Entamoeba histolytica* Genome. In *Advances in Parasitology*; Elsevier: Amsterdam, The Netherlands, 2007; Volume 65, pp. 51–190. ISBN 978-0-12-374166-0.
84. Lorenzi, H.A.; Puiu, D.; Miller, J.R.; Brinkac, L.M.; Amedeo, P.; Hall, N.; Caler, E.V. New Assembly, Reannotation and Analysis of the *Entamoeba histolytica* Genome Reveal New Genomic Features and Protein Content Information. *PLoS Negl. Trop. Dis.* **2010**, *4*, e716. [CrossRef]
85. Weedall, G.D.; Hall, N. Evolutionary Genomics of *Entamoeba*. *Res. Microbiol.* **2011**, *162*, 637–645. [CrossRef] [PubMed]
86. Naiyer, S.; Kaur, D.; Ahamad, J.; Singh, S.S.; Singh, Y.P.; Thakur, V.; Bhattacharya, A.; Bhattacharya, S. Transcriptomic Analysis Reveals Novel Downstream Regulatory Motifs and Highly Transcribed Virulence Factor Genes of *Entamoeba histolytica*. *BMC Genom.* **2019**, *20*, 206. [CrossRef]
87. Bhattacharya, S. Episomal and Chromosomal DNA Replication and Recombination in *Entamoeba histolytica*. *Front. Mol. Biosci.* **2023**, *10*, 1212082. [CrossRef]
88. Singh, N.; Bhattacharya, A.; Bhattacharya, S. Homologous Recombination Occurs in *Entamoeba* and Is Enhanced during Growth Stress and Stage Conversion. *PLoS ONE* **2013**, *8*, e74465. [CrossRef]
89. Romero, M.; Cerritos, R.; Ximenez, C. Horizontal Gene Transfers from Bacteria to *Entamoeba* Complex: A Strategy for Dating Events along Species Divergence. *J. Parasitol. Res.* **2016**, *2016*, 3241027. [CrossRef] [PubMed]
90. Gilchrist, C.A.; Houpt, E.; Trapaidze, N.; Fei, Z.; Crasta, O.; Asgharpour, A.; Evans, C.; Martino-Catt, S.; Baba, D.J.; Stroup, S.; et al. Impact of Intestinal Colonization and Invasion on the *Entamoeba histolytica* Transcriptome. *Mol. Biochem. Parasitol.* **2006**, *147*, 163–176. [CrossRef]
91. Davis, P.H.; Schulze, J.; Stanley, S.L. Transcriptomic Comparison of Two *Entamoeba histolytica* Strains with Defined Virulence Phenotypes Identifies New Virulence Factor Candidates and Key Differences in the Expression Patterns of Cysteine Proteases, Lectin Light Chains, and Calmodulin. *Mol. Biochem. Parasitol.* **2007**, *151*, 118–128. [CrossRef]
92. Davis, P.H.; Zhang, Z.; Chen, M.; Zhang, X.; Chakraborty, S.; Stanley, S.L. Identification of a Family of BspA like Surface Proteins of *Entamoeba histolytica* with Novel Leucine Rich Repeats. *Mol. Biochem. Parasitol.* **2006**, *145*, 111–116. [CrossRef]
93. König, C.; Honecker, B.; Wilson, I.W.; Weedall, G.D.; Hall, N.; Roeder, T.; Metwally, N.G.; Bruchhaus, I. Taxon-Specific Proteins of the Pathogenic *Entamoeba* Species *E. histolytica* and *E. nuttalli*. *Front. Cell. Infect. Microbiol.* **2021**, *11*, 641472. [CrossRef]
94. Das, K.; Ganguly, S. Evolutionary Genomics and Population Structure of *Entamoeba histolytica*. *Comput. Struct. Biotechnol. J.* **2014**, *12*, 26–33. [CrossRef] [PubMed]
95. Kumari, V.; Iyer, L.R.; Roy, R.; Bhargava, V.; Panda, S.; Paul, J.; Verweij, J.J.; Clark, C.G.; Bhattacharya, A.; Bhattacharya, S. Genomic Distribution of SINEs in *Entamoeba histolytica* Strains: Implication for Genotyping. *BMC Genom.* **2013**, *14*, 432. [CrossRef]
96. Urquieta-Ramírez, L.; Ramírez-Montiel, F.; Andrade-Guillén, S.; Páramo-Pérez, I.; Rangel-Serrano, Á.; Reyes-Cortes, R.; Franco, B.; Mendoza-Macías, C.L.; Anaya-Velázquez, F.; Padilla-Vaca, F. Contribution of Neutral Sphingomyelinases to in Vitro Virulence of *Entamoeba histolytica*. *Exp. Parasitol.* **2018**, *194*, 38–44. [CrossRef]
97. Wilson, I.W.; Weedall, G.D.; Lorenzi, H.; Howcroft, T.; Hon, C.-C.; Deloger, M.; Guillén, N.; Paterson, S.; Clark, C.G.; Hall, N. Genetic Diversity and Gene Family Expansions in Members of the Genus *Entamoeba*. *Genome Biol. Evol.* **2019**, *11*, 688–705. [CrossRef]
98. Nakada-Tsukui, K.; Sekizuka, T.; Sato-Ebine, E.; Escueta-de Cadiz, A.; Ji, D.; Tomii, K.; Kuroda, M.; Nozaki, T. AIG1 Affects in Vitro and in Vivo Virulence in Clinical Isolates of *Entamoeba histolytica*. *PLoS Pathog.* **2018**, *14*, e1006882. [CrossRef]
99. Tabernero, L.; Aricescu, A.R.; Jones, E.Y.; Szedlacsek, S.E. Protein Tyrosine Phosphatases: Structure-Function Relationships: PTP Structure-Function Relationship. *FEBS J.* **2008**, *275*, 867–882. [CrossRef] [PubMed]
100. Linford, A.S.; Jiang, N.M.; Edwards, T.E.; Sherman, N.E.; Van Voorhis, W.C.; Stewart, L.J.; Myler, P.J.; Staker, B.L.; Petri, W.A. Crystal Structure and Putative Substrate Identification for the *Entamoeba histolytica* Low Molecular Weight Tyrosine Phosphatase. *Mol. Biochem. Parasitol.* **2014**, *193*, 33–44. [CrossRef] [PubMed]
101. Torres-Cifuentes, D.M.; Espíritu-Gordillo, P.; Baylón-Pacheco, L.; Rosales-Encína, J.L. Low Molecular Weight Protein Tyrosine Phosphatase (LMW-PTP2) Protein Can Potentially Modulate Virulence of the Parasite *Entamoeba histolytica*. *Mol. Biochem. Parasitol.* **2021**, *242*, 111360. [CrossRef] [PubMed]
102. Gastelum-Martínez, A.; León-Sicairos, C.; Plata-Guzmán, L.; Soto-Castro, L.; León-Sicairos, N.; De La Garza, M. Iron-Modulated Virulence Factors of *Entamoeba histolytica*. *Future Microbiol.* **2018**, *13*, 1329–1341. [CrossRef] [PubMed]
103. Rath, P.P.; Gourinath, S. The Actin Cytoskeleton Orchestra in *Entamoeba histolytica*. *Proteins* **2020**, *88*, 1361–1375. [CrossRef]
104. Medina-Rosales, M.N.; Muñoz-Ortega, M.H.; García-Hernández, M.H.; Talamás-Rohana, P.; Medina-Ramírez, I.E.; Salas-Morón, L.G.; Martínez-Hernández, S.L.; Ávila-Blanco, M.E.; Medina-Rosales, B.; Ventura-Juárez, J. Acetylcholine Upregulates *Entamoeba histolytica* Virulence Factors, Enhancing Parasite Pathogenicity in Experimental Liver Amebiasis. *Front. Cell. Infect. Microbiol.* **2021**, *10*, 586354. [CrossRef]
105. Leon-Coria, A.; Kumar, M.; Chadee, K. The Delicate Balance between *Entamoeba histolytica*, Mucus and Microbiota. *Gut Microbes.* **2020**, *11*, 118–125. [CrossRef]
106. Iyer, L.R.; Verma, A.K.; Paul, J.; Bhattacharya, A. Phagocytosis of Gut Bacteria by *Entamoeba histolytica*. *Front. Cell. Infect. Microbiol.* **2019**, *9*, 34. [CrossRef]

107. Partida-Rodríguez, O.; Serrano-Vázquez, A.; Nieves-Ramírez, M.E.; Moran, P.; Rojas, L.; Portillo, T.; González, E.; Hernández, E.; Finlay, B.B.; Ximenez, C. Human Intestinal Microbiota: Interaction between Parasites and the Host Immune Response. *Arch. Med. Res.* **2017**, *48*, 690–700. [CrossRef]
108. Burgess, S.L.; Petri, W.A. The Intestinal Bacterial Microbiome and *E. histolytica* Infection. *Curr. Trop. Med. Rep.* **2016**, *3*, 71–74. [CrossRef]
109. Guillén, N. The Interaction between *Entamoeba histolytica* and Enterobacteria Shed Light on an Ancient Antibacterial Response. *Cell. Microbiol.* **2019**, *21*, 13039. [CrossRef] [PubMed]
110. Varet, H.; Shaulov, Y.; Sismeiro, O.; Trebicz-Geffen, M.; Legendre, R.; Coppée, J.-Y.; Ankri, S.; Guillen, N. Enteric Bacteria Boost Defences against Oxidative Stress in *Entamoeba histolytica*. *Sci. Rep.* **2018**, *8*, 9042. [CrossRef] [PubMed]
111. Fernández-López, L.A.; Gil-Becerril, K.; Galindo-Gómez, S.; Estrada-García, T.; Ximénez, C.; Leon-Coria, A.; Moreau, F.; Chadee, K.; Tsutsumi, V. *Entamoeba histolytica* Interaction with Enteropathogenic *Escherichia coli* Increases Parasite Virulence and Inflammation in Amebiasis. *Infect. Immun.* **2019**, *87*, e00279-19. [CrossRef] [PubMed]
112. Silva Oliveira, F.M.; Claúdia De Carvalho Fernandes, A.; Henrique De Cicco Sandes, S.; Prata, L.O.; Junior, M.A.; Vilela Da Silva, C.A.; Campolina-Silva, G.H.; Lorenzo De Jesus Oliveira, M.; Oliveira, C.A.; Neumann, E.; et al. Co-Infection by *Salmonella enterica* Subsp. Enterica Serovar Typhimurium and *Entamoeba dispar* Pathogenic Strains Enhances Colitis and the Expression of Amoebic Virulence Factors. *Microb. Pathog.* **2021**, *158*, 105010. [CrossRef]
113. Verma, A.K.; Verma, R.; Ahuja, V.; Paul, J. Real-Time Analysis of Gut Flora in *Entamoeba histolytica* Infected Patients of Northern India. *BMC Microbiol.* **2012**, *12*, 183. [CrossRef]
114. Morton, E.R.; Lynch, J.; Froment, A.; Lafosse, S.; Heyer, E.; Przeworski, M.; Blekhman, R.; Ségurel, L. Variation in Rural African Gut Microbiota Is Strongly Correlated with Colonization by *Entamoeba* and Subsistence. *PLoS Genet.* **2015**, *11*, e1005658. [CrossRef]
115. Serrano-Luna, J.; Piña-Vázquez, C.; Reyes-López, M.; Ortiz-Estrada, G.; De La Garza, M. Proteases from *Entamoeba* spp. and Pathogenic Free-Living Amoebae as Virulence Factors. *J. Trop. Med.* **2013**, *2013*, 1–32. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

LETTER

doi:10.1038/nature10553

Ascaris suum draft genome

Aaron R. Jex[1]*, Shiping Liu[2]*, Bo Li[2]*, Neil D. Young[1]*, Ross S. Hall[1], Yingrui Li[2], Linfeng Yang[2], Na Zeng[2], Xun Xu[2], Zijun Xiong[2], Fangyuan Chen[2], Xuan Wu[2], Guojie Zhang[2], Xiaodong Fang[2], Yi Kang[2], Garry A. Anderson[1], Todd W. Harris[3], Bronwyn E. Campbell[1], Johnny Vlaminck[4], Tao Wang[4], Cinzia Cantacessi[1], Erich M. Schwarz[5], Shoba Ranganathan[6], Peter Geldhof[4], Peter Nejsum[7], Paul W. Sternberg[5], Huanming Yang[2], Jun Wang[2], Jian Wang[2] & Robin B. Gasser[1]

Parasitic diseases have a devastating, long-term impact on human health, welfare and food production worldwide. More than two billion people are infected with geohelminths, including the roundworms *Ascaris* (common roundworm), *Necator* and *Ancylostoma* (hookworms), and *Trichuris* (whipworm), mainly in developing or impoverished nations of Asia, Africa and Latin America[1]. In humans, the diseases caused by these parasites result in about 135,000 deaths annually, with a global burden comparable with that of malaria or tuberculosis in disability-adjusted life years[1]. *Ascaris* alone infects around 1.2 billion people and, in children, causes nutritional deficiency, impaired physical and cognitive development and, in severe cases, death[2]. *Ascaris* also causes major production losses in pigs owing to reduced growth, failure to thrive and mortality[2]. The *Ascaris*–swine model makes it possible to study the parasite, its relationship with the host, and ascariasis at the molecular level. To enable such molecular studies, we report the 273 megabase draft genome of *Ascaris suum* and compare it with other nematode genomes. This genome has low repeat content (4.4%) and encodes about 18,500 protein-coding genes. Notably, the *A. suum* secretome (about 750 molecules) is rich in peptidases linked to the penetration and degradation of host tissues, and an assemblage of molecules likely to modulate or evade host immune responses. This genome provides a comprehensive resource to the scientific community and underpins the development of new and urgently needed interventions (drugs, vaccines and diagnostic tests) against ascariasis and other nematodiases.

We sequenced the *A. suum* genome at ~80-fold coverage (Supplementary Fig. 1), producing a final draft assembly of 272,782,664 base pairs (bp) (N50 = 407 kilobases, kb; N90 = 80 kb; 1,618 contigs of >2 kb) (Table 1) with a mean GC-content of 37.9%. This genome has few repetitive sequences (about 4.4% of the total assembly) relative to that reported for other metazoan genomes sequenced to date[3–6], probably as a result of chromatin diminution[7]. We identified 424 distinct retrotransposon sequences (see Supplementary Tables 1–3) representing at least 22 families (8 long terminal repeats (LTRs), 12 long interspersed elements (LINEs) and 2 short interspersed elements (SINEs)), with *Gypsy*, *Pao* and *Copia* classes predominating for LTRs ($n = 97$, 85 and 60, respectively) and CR1, L1, and reverse transcriptase encoding RTE-RTE classes predominating for non-LTRs ($n = 29$, 28 and 21, respectively). We also identified eight families of DNA transposons (91 distinct sequences in total), of which *MuDr*, *En-Spm* and *Merlin* ($n = 12$, 9 and 8, respectively) predominated. We predicted 18,542 genes (14,783 supported by transcriptomic data), with a mean total length of 6.5 kb, exon length of 153 bp and a mean of 6.4 exons per gene (see Supplementary Fig. 2). Compared with the nematodes (roundworms) *Caenorhabditis elegans*[3], *Pristionchus pacificus*[8], *Brugia malayi*[9] or *Meloidogyne hapla*[10], overall, the *A. suum* genes are significantly longer (see Supplementary Table 2), relating primarily to expansions of intronic regions (mean 1.1 kb).

Table 1 | Features of the *Ascaris suum* draft genome

Estimated genome size in megabases	309
Total number of base pairs within assembled scaffolds	272,782,664
N50 length in bp; total number >2 kb in length	407,899; 1,618
N90 length in bp; total number >N90 length	80,017; 748
GC content of whole genome (%)	37.9
Repetitive sequences (%)	4.4
Proportion of genome that is coding (exonic; including introns) (%)	5.9; 44.2
Number of putative coding genes	18,542
Gene size (mean bp)	6,536
Average coding domain length (mean bp)	983
Average exon number per gene (mean)	6
Gene exon length (mean bp)	153
Gene intron length (mean bp)	1,081
GC content in coding regions (%)	45
Number of transfer RNAs	255

N50 means 50% of all nucleotides in the assembly are within contigs of ≥408 kb. N90 means 90% of all nucleotides in the assembly are within contigs of ≥80 kb. Genome size estimated on the basis of *k*-mer (see online-only Methods) frequency.

Most (78.2%) of the predicted *A. suum* genes (Fig. 1) have a homologue (BLASTp cut-off $\leq 10^{-5}$) either in *C. elegans* ($n = 12{,}779$; 68.9%), *B. malayi* (12,853; 69.3%), *M. hapla* (10,482; 56.5%) or *P. pacificus* (11,865; 64.0%), with 8,967 being homologous among all species examined, and 4,042 (21.8%) being 'unique' to *A. suum* (see Fig. 1). Of the genes with homology to *C. elegans* or *B. malayi*, ~50%

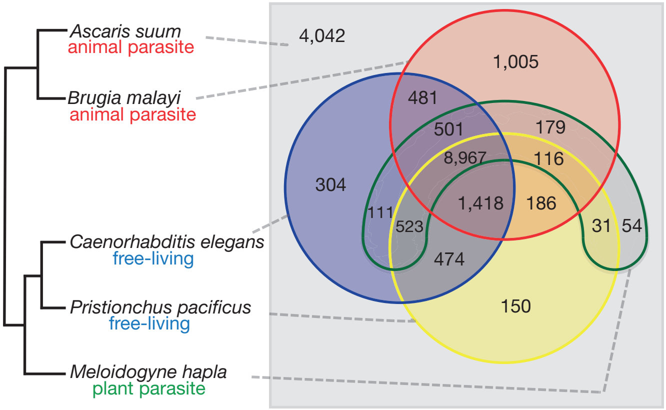


Figure 1 | Venn diagram summarizing the overlapping homology between the *Ascaris suum* gene set and those of other nematodes. Grey box (right) represents genes unique to *A. suum*, relative to *Brugia malayi* (red circle), *Caenorhabditis elegans* (blue circle), *Meloidogyne hapla* (green arc) and/or *Pristionchus pacificus* (yellow circle). The phylogram (left) displays the evolutionary relationships currently proposed among the nematodes.

---

[1]Faculty of Veterinary Science, The University of Melbourne, Parkville, Victoria 3010, Australia. [2]BGI-Shenzhen, Shenzhen, 518083, China. [3]Ontario Institute for Cancer Research, MaRS Centre, South Tower, 101 College Street, Suite 800, Toronto, Ontario, M5G 0A3, Canada. [4]Laboratory of Parasitology and Parasitic Diseases, University of Ghent, Merelbeke B-9820, Belgium. [5]California Institute of Technology, Pasadena, California, 91125, USA. [6]Department of Chemistry and Biomolecular Sciences, Macquarie University, Sydney, New South Wales 2109, Australia. [7]Faculty of Life Sciences, University of Copenhagen, Copenhagen DK-2200, Denmark.

*These authors contributed equally to this work.

©2011 Macmillan Publishers Limited. All rights reserved

and 44%, respectively, were determined to represent one-to-one orthologues[11] (see Supplementary Data 1). For these orthologues (on scaffolds exceeding one megabase, 1 Mb, in size), we explored synteny for *A. suum* and *B. malayi* by pairwise comparison with *C. elegans* (see Supplementary Data 1). The findings show that interchromosomal gene rearrangements in *A. suum* are relatively rare and occurred less frequently in *A. suum* than in *B. malayi*[9] relative to *C. elegans* since their evolutionary divergence[12]. In contrast, intrachromosomal rearrangements were relatively common and comparable in frequency to those inferred for *B. malayi*[9]. Overall synteny was significantly higher between *A. suum* and *B. malayi* (∼15%) than between either species and *C. elegans* (∼3%), which is consistent with current knowledge of the evolutionary relationships among these three species[12]. Interestingly, of these *C. elegans* orthologous genes, 532 and 483 were exclusive to the current assemblies of the *A. suum* and *B. malayi* genomes, respectively (Supplementary Data 2). Although there were no homology matches between these two exclusive subsets of orthologues, they shared striking similarity in functional ontology (biological process), being linked predominantly to growth, reproduction, development and/or morphogenesis. There is clear evidence of plasticity in the germline of metazoans[13], with cases of products from non-homologous genes in different species having analogous function(s). Therefore, we hypothesize that these two unique gene subsets relate to differences in reproductive biology (oviparity versus viviparity) and life history (direct versus indirect) between *A. suum* and *B. malayi*. Clearly, this proposal warrants testing and functional validation in *C. elegans* and/or in *Ascaris*.

Of the entire *A. suum* gene set, 2,370 genes had an orthologue (BLASTp cut-off $\leq 10^{-5}$) belonging to one of 279 known biological (KEGG; see online-only Methods) pathways (Supplementary Data 3). Mapping to pathways in *C. elegans* indicated a full complement of molecules; by inference, the vast majority (95%) of the *A. suum* euchromatin is represented in the present genomic assembly, an inference that is supported by our transcriptomic data (Supplementary Tables 4 and 5). We were able to assign possible functions (such as for enzymes, receptors, channels and transporters; Supplementary Fig. 3, Supplementary Table 6 and Supplementary Data 4) to 13,503 (72.8%) of the genes predicted for *A. suum* (Fig. 2). For these genes, we predicted 456 peptidases belonging to five major classes (aspartic, cysteine, metallo-, serine and threonine), with the metallo- ($n = 184$: 41.0%) and serine proteases ($n = 132$: 30.0%) predominating (Supplementary Data 4). Notably, the secreted peptidases (such as the M12 'astacins', the S9 and S33 serine proteases, and the C1 and C2 cysteine proteases) are abundantly represented, and have key roles in tissue invasion and degradation (for example, during migration and/or feeding) and/or immune evasion/modulation in many parasites[14,15].

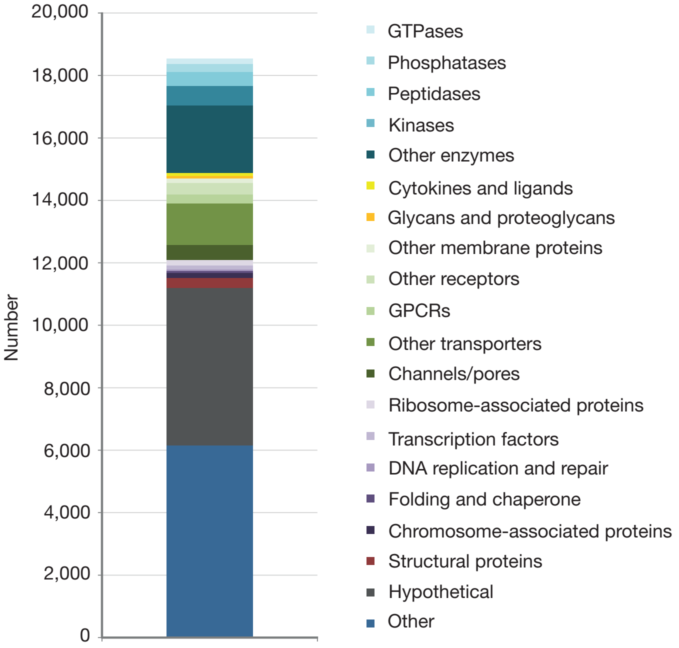


Figure 2 | The major protein classes representing the *Ascaris suum* gene set.

In addition, we identified 609 kinases and 257 phosphatases, respectively (Supplementary Data 4). All major classes of kinases are represented, with the tyrosine (TK: $n = 94$), casein (CK1: $n = 83$), CMGC ($n = 67$) and CAMK ($n = 54$) being most abundant in *A. suum*. The phosphatome includes 17 receptor and 68 conventional tyrosine, 64 serine/threonine and 39 dual-specificity phosphatases. On the basis of homology with molecules in *C. elegans*, 169 GTPases are encoded in the *A. suum* genome, including 135 small GTPases (Ras superfamily) representing the Rab ($n = 36$), Ras ($n = 35$; plus 8 Ras-like), Rho ($n = 17$; plus 9 Rho-like) or Ran ($n = 6$) subfamilies. Examples of these homologues include *eft-1*, *fzo-1*, *glo-1* and *rho-1*, which have essential roles in embryonic, larval and/or reproductive development (see www.wormbase.org).

Given their key roles, many of these enzymes are proposed as targets for anti-parasitic compounds and/or vaccines[16–18]. Equally, the range of receptor and channel proteins identified here are interesting because many common anthelmintics bind such targets[19]. Here, we predicted 279 G protein-coupled receptors (GPCRs) for *A. suum* and 477 channel or pore proteins (Supplementary Data 4), including 272 voltage-gated and 98 ligand-gated ion channels. Many voltage-gated ion channels are known targets for nematocidal drugs, such as macrocyclic lactones (for example, ivermectin) and levamisole, and an aminoacetonitrile derivative, monepantel, is the most recent example of a highly effective nematocide that binds to a ligand-gated ion channel[19]. Importantly, in the *A. suum* gene set, we found a homologue (*acr-23*) of the *C. elegans* monepantel receptor[19], suggesting that this drug may kill *A. suum*. In addition, we detected 462 transporters (for example, small molecule porter proteins), of which the major facilitator ($n = 155$), cation symporter ($n = 71$) and resistance-nodulation-cell division ($n = 56$) superfamilies were most abundant (Supplementary Data 4).

Excretory/secretory (E/S) peptides are central to understanding parasite–host interactions. We predicted the secretome of *A. suum* to comprise 775 proteins with diverse functions (Supplementary Data 5). Notable among them are 68 secreted proteases, including 20 SC clan serine proteases (S9 and S33 families), 18 MA clan metallo-proteases (M10, M12 and M41 families) and 5 CA/CD clan cysteine proteases (C1 and C13 families); see http://merops.sanger.ac.uk/ for clan definitions.

Secreted proteases have known roles in host-tissue degradation, required for feeding, tissue-penetration and/or larval migration for a range of helminths[14], including *Ascaris*[2]. In addition, they are involved in inducing and modulating host immune responses against parasitic helminths[15], which are often Th2-biased[20]. From the current understanding of these responses[15], we compiled a comprehensive list of *A. suum* E/S proteins homologous to helminth-secreted peptides with important immunogenic or immunomodulatory roles in host animals (Supplementary Table 7 and Supplementary Data 6). Such homologues represent about half of the predicted *A. suum* secretome. Most abundant among them are O-linked glycosylated proteins ($n = 300$), many of which are heavily targeted by immunoglobulin (Ig) M antibodies and bound by various pattern recognition receptors associated with host dendritic cells responsible for the induction of a Th2 immune response[15].

Other members of the *A. suum* secretome are predicted to direct or evade immune responses. These peptides include a close homologue of the E/S-62 leucyl aminopeptidase of the filarioid nematode *Acanthocheilonema viteae*, which has been shown to inhibit B-cell, T-cell and mast cell proliferation/responses, promote an alternative activation of the host macrophages, through the inhibition of the Toll-like receptor signalling pathway, and induce a Th2 response through

©2011 Macmillan Publishers Limited. All rights reserved

the inhibition of IL-12p70 production by dendritic cells[15]. Additional, immunomodulatory molecules predicted for *A. suum* (BLASTp cut-off $\leq 10^{-5}$) include homologues of another B-cell inhibitor (that is, the *B. malayi* cystatin CPI-2), several TGF-β and macrophage initiation factor mimics, numerous neutrophil inhibitors, various oxidoreductases, and five close homologues of platelet anti-inflammatory factor α (ref. 15). Some *A. suum* E/S peptides are predicted to be involved in immune evasion; for instance, some mask parasite antigens by mimicking host molecules (such as several C-type lectins with close homology to vertebrate macrophage mannose or CD23 (low affinity IgE receptors[15]).

Taken together, these data indicate that *A. suum* has a large arsenal of E/S proteins that are likely to be involved directly in manipulating, blocking and/or evading immune responses in the host. Understanding the immunomolecular interplay between *A. suum* and its host, early in infection, particularly during hepatopulmonary migration, should pave the way for designing prophylactic interventions, such as vaccination.

Ascaris larvae undertake an extensive migration through their host's body before they establish as adults in the small intestine. Following the ingestion of infective eggs and their gastric passage, third-stage larvae (L3s)[21] hatch from eggs in the gut and penetrate the intestinal wall; they then undergo, via the bloodstream, an arduous hepatopulmonary migration. The complexity of this migration coincides with important developmental changes in the nematode[2]. Clearly, this migration requires tightly regulated transcriptional changes in the parasite. We explored this aspect by characterizing the transcription profiles of infective L3s (from eggs), L3s from the liver or lungs of the host, and fourth-stage larvae (L4s) from the small intestine (Supplementary Fig. 4, Supplementary Data 7). Notable among genes enriched during larval migration are various secreted peptidases linked to tissue-penetration and degradation during feeding and/or migration[14], including three C1/C2, five M1, eight M12, fourteen S9 and five S33 clan members. Considering the complex nature of larval migration, a key role for molecules associated with chemosensory pathways is highly likely. Such molecules have been studied extensively in *C. elegans*[22], with numerous homologues being identified here in larval transcripts (Supplementary Data 7). With few exceptions, all of these homologues relate to olfactory chemosensation of volatile compounds (for example, alcohols, aldehydes or ketones), suggesting that the olfactory detection of molecular gradients is central to the navigation of *A. suum* larvae during migration. Lastly, considering the substantial host attack against migrating *Ascaris* larvae, E/S proteins probably play crucial roles in immune modulation and/or evasion during hepatopulmonary migration. Many such genes, including *Bm-alt-1*, *Bm-cpi-2* and *mif-4*, are highly transcribed in *A. suum* larvae (see Supplementary Data 7), particularly in migrating L3s.

Because of the large size of the adult nematode (10–15 cm), we were able to explore transcription in the musculature and reproductive tracts of adult male and female *A. suum* individuals as well (Supplementary Fig. 5 and Supplementary Data 8). Among the male-enriched transcripts is a range of genes associated specifically with sperm and/or spermatogenesis, including *fer-1*, *spe-4*, *spe-6*, *spe-9*, *spe-10*, *spe-15* and *spe-41*, *alg-4* and *msp-57* (see www.wormbase.org). Notable among the female-enriched transcripts is a large variety of genes associated with oogenesis/egg-laying (such as *cat-1*, *unc-54*, *cbd-1* and *pqn-74*), vulval development (such as *noah-1*, *nhr-25*, *cog-1* and *pax-3*) and/or embryogenesis (such as *cam-1* and *unc-6*; see www.wormbase.org). Although the functions of these genes have been explored in *C. elegans* (primarily a hermaphroditic nematode), this detailed insight into the tissue-specific transcription for a dioecious nematode is a major advance.

Analyses of these RNA-seq data revealed 163,777 single nucleotide polymorphisms (SNPs) in coding regions of the *A. suum* genome; 61% of them were synonymous, 7% non-synonymous and <0.1% termination codons (Supplementary Data 9). Some of the most variable genes in *A. suum* encoded ribosomal proteins ($n = 44$), translation initiation factor (*tif*) eIF-3 subunits 3 and 5, *tif* TFIIH subunit H2 and *tif* IF-2, galectin-4 and galectin-9, the latter two of which are probably linked to immune evasion[15] and may indicate that antigenic variation is among the many strategies apparent in *Ascaris* to combat the host immune response. Interestingly, the high nucleotide variability linked to the key elements of translation machinery did not relate to a bias in synonymous SNPs, suggesting that many mutations accumulate in particular 'hotspots' and/or are tolerated, but do not compromise either the structure or the function of this machinery. The least variable genes encoded various (druggable)[16,17] serine/threonine phosphatases ($n = 17$) as well as numerous receptors, channels and transporters, for which there was an unusually strong bias towards synonymous SNPs, reinforcing their potential as intervention targets.

Given our present reliance on a small number of drugs (for example, piperazine, pyrantel, albendazole and mebendazole) for the treatment of ascariasis, their repeated or excessive use might lead to resistance in *Ascaris* populations to some or all of these compounds[23]. As few new anthelmintics (that is, aminoacetylnitriles[19] and cyclooctodepsipeptides[24]) have been discovered in the past two decades using traditional screening methods, an effective, alternative means of drug discovery is urgently needed[23]. Genome-guided drug target or drug discovery has major potential to complement conventional screening and re-purposing. The goal of genome-guided analysis is to identify genes or molecules whose inactivation by one or more drugs will selectively kill parasites but not harm their host.

Because most parasitic nematodes are difficult to produce or maintain outside of their host, or to subject to gene-specific silencing by RNAi[23] or morpholinos[25,26], direct functional assessment of essentiality (that is, they are needed for the nematode's survival) is not yet practical. However, essentiality can be inferred from functional information for model organisms (for example, lethality in *C. elegans* and *D. melanogaster*)[27], and this approach has indeed yielded effective targets for nematocides[16]. In *Ascaris*, we identified 629 proteins (Supplementary Data 10) with essential homologues in *C. elegans* and *D. melanogaster* (linked to lethal phenotypes upon gene perturbation). Among these are 87 channels or transporters (including 44 voltage-gated ion channels), which represent protein classes most successfully targeted for anthelmintic compounds, including macrocyclic lactones, levamisoles and aminoacetonitrile

Table 2 | Druggable candidates in the *Ascaris suum* draft genome

Protein or chokepoint	Subtype (number of molecules)	Total number
GTPase	Small GTPase (22); Ras (13); Rab (5); Rho (3); Ras-like (1); others (2)	46
Kinase	TK (8); AGC (3); CAMK (2); TKL (2); STE (1); other (1)	17
Peptidase	A22A (5); M14B (3); M12B (2); M67A (1); C14A (1); C50 (1); M12A (1); M13 (1); T01A (1); C46 (1); S33 (1)	19
Phosphatase	STP (28); cPTP (4); DSP (3)	35
Transporters and channels	Channels and pores (30); primary active transporters (24); incompletely characterized transport system proteins (22); accessory factors involved in transport (5); electrochemical potential-driven transporters (5); group translocators (1)	87
'Lethal' chokepoints	CDP-diacylglycerol-inositol 3-phosphatidyltransferase	1
	G protein-coupled receptor kinase 5	1
	Phosphoribosylformylglycinamidine synthase	1
	Inosine-5′-monophosphate dehydrogenase	1
	Phospho-*N*-acetylmuramoyl pentapeptide transferase	1

Candidates were inferred from essentiality prediction and metabolic chokepoint analysis.

©2011 Macmillan Publishers Limited. All rights reserved

derivatives[19,28]. Also notable are 46 GTPases, 35 phosphatases (including PP1 and PP2A homologues, as targets for norcantharidin analogues)[16], 17 kinases and 19 peptidases (Table 2).

In addition to essentiality-based prediction, an alternative strategy has been to infer enzymatic chokepoints intrinsic to the complete metabolome of a parasite[29]. Such chokepoints are defined as enzymatic reactions that uniquely produce and/or consume a molecular compound, using the strategy that the disruption of such enzymes would lead to the toxic build-up (that is, for unique substrates) or starvation (that is, for unique products) of metabolites within cells. Pathway analysis identified 225 likely chokepoints linked to genes predicted to be essential in *A. suum* (Supplementary Data 10). We gave the highest priority to targets predicted from single-copy genes in the *A. suum* genome, reasoning that lower allelic variability would exist within populations and would thus be less likely to give rise to drug resistance.

Using this strategy, we identified five high-priority drug targets for *A. suum* (see Table 2 and Supplementary Data 10) that, given their conservation with *C. elegans* and *D. melanogaster*, are likely to be relevant in relation to many other parasitic helminths. Conspicuous among them is IMP dehydrogenase (GMP reductase), which has a variety of inhibitors (for example, mycophenolic acid analogues[30]) that could be tested for ascaricidal effects. Clearly, the druggable genome of *Ascaris* now provides a solid basis for rational drug design, aimed at controlling parasitic nematodes of major socioeconomic impact worldwide.

In conclusion, we have characterized the genome of *A. suum*, a major parasite of one of the world's most important food animals (pig) and the closest relative of *A. lumbricoides*, which infects about 1.2 billion people globally[1,2]. Intriguingly, the present *A. suum* draft genome exhibits unusually low repeat content and lacks Tas2 transposons[7]. These characteristics probably relate to the chromatin diminution described previously for some ascaridoids[7], indicating that our assembly represents the somatic genome of this parasite. The precise mechanism governing this diminution is not yet understood. Although the chromatin lost during this process is not fully characterized, there appears to be a significant loss in repeat content[7], consistent with the present assembly. Notably, the present gene set inferred for *A. suum* includes *fert-1* and *rpS19G*, which, although originally proposed to be germline-specific[7], were transcribed in all adult libraries sequenced here. This finding suggests that the genomic content lost during diminution might vary among individuals or tissues, and is a stimulus to investigate chromatin diminution between and among individual cells (that is, sperm or eggs), stages and tissue types of *A. suum*. Importantly, the present study, showing that a high-quality genomic assembly can be achieved using an approach based on whole-genome amplification, provides unique prospects for exploring diminution in detail, using the present genome as a reference.

In addition, our sequencing effort has characterized a broad range of key classes of molecules of major relevance to understanding the molecular biology of *A. suum* and the exquisite complexities of the host–parasite interplay on an immunobiological level. This work paves the way for future fundamental molecular explorations and the design of new methods for the treatment and control of one of the world's most important parasitic nematodes. This focus is now crucial, given the major impact of *Ascaris* and other soil-transmitted helminths, which affect billions of people and animals worldwide. Although these parasites are seriously neglected, genomic and post-genomic approaches provide new hope for the discovery of intervention strategies, with major implications for improving global health.

METHODS SUMMARY

We sequenced the genome of *A. suum* using Illumina technology from genomic DNA from the reproductive tract of a single adult female. From six paired-end sequencing libraries (insert sizes: 0.17 kb to 10 kb; see Supplementary Tables 1 and 2), we generated 39 Gb of useable short-read sequence data, equating to ~80-fold coverage of the 273-Mb genome. We assembled the short reads, constructed scaffolds in a step-by-step manner, and then closed intra-scaffold gaps[5]. Transposable elements, non-coding RNAs and the protein-coding gene set were inferred using a combination of predictive modelling and a homology-based approach. Orthology and synteny analyses were conducted using established methods[9,11]. We sequenced messenger RNA from infective L3s (from eggs), migrating L3s from the liver or lungs of the host, and L4s from the small intestine, as well as muscle and reproductive tissues from adult male and female *A. suum*, and used these data to aid gene predictions, define SNPs and explore key molecules associated with larval migration, reproduction and development. All proteins predicted from the gene set were annotated using databases for conserved protein domains, gene ontology annotations and model organisms (that is, *Caenorhabditis elegans*, *Drosophila melanogaster* and *Mus musculus*). Essentiality and drug target predictions were conducted using established or in-house methods.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 16 June; accepted 12 September 2011.

Published online 26 October 2011.

1. Hotez, P. J., Fenwick, A., Savioli, L. & Molyneux, D. H. Rescuing the bottom billion through control of neglected tropical diseases. *Lancet* **373,** 1570–1575 (2009).
2. Crompton, D. W. *Ascaris* and ascariasis. *Adv. Parasitol.* **48,** 285–375 (2001).
3. The *C. elegans* Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282,** 2012–2018 (1998).
4. Adams, M. D. *et al.* The genome sequence of *Drosophila melanogaster*. *Science* **287,** 2185–2195 (2000).
5. Li, R. *et al.* The sequence and *de novo* assembly of the giant panda genome. *Nature* **463,** 311–317 (2010).
6. Mitreva, M. *et al.* The draft genome of the parasitic nematode *Trichinella spiralis*. *Nature Genet.* **43,** 228–235 (2011).
7. Müller, F. & Tobler, H. Chromatin diminution in the parasitic nematodes *Ascaris suum* and *Parascaris univalens*. *Int. J. Parasitol.* **30,** 391–399 (2000).
8. Dieterich, C. *et al.* The *Pristionchus pacificus* genome provides a unique perspective on nematode lifestyle and parasitism. *Nature Genet.* **40,** 1193–1198 (2008).
9. Ghedin, E. *et al.* Draft genome of the filarial nematode parasite *Brugia malayi*. *Science* **317,** 1756–1760 (2007).
10. Opperman, C. H. *et al.* Sequence and genetic map of *Meloidogyne hapla*: a compact nematode genome for plant parasitism. *Proc. Natl Acad. Sci. USA* **105,** 14802–14807 (2008).
11. Kuzniar, A., van Ham, R. C. H. J., Pongor, S. & Leunissen, J. A. M. The quest for orthologs: finding the corresponding gene across genomes. *Trends Genet.* **24,** 539–551 (2008).
12. Blaxter, M. L. *et al.* A molecular evolutionary framework for the phylum Nematoda. *Nature* **392,** 71–75 (1998).
13. Ewen-Campen, B., Schwager, E. E. & Extavour, C. G. The molecular machinery of germ line specification. *Mol. Reprod. Dev.* **77,** 3–18 (2010).
14. McKerrow, J. H., Caffrey, C., Kelly, B., Loke, P. & Sajid, M. Proteases in parasitic diseases. *Annu. Rev. Pathol.* **1,** 497–536 (2006).
15. Hewitson, J. P., Grainger, J. R. & Maizels, R. M. Helminth immunoregulation: the role of parasite secreted proteins in modulating host immunity. *Mol. Biochem. Parasitol.* **167,** 1–11 (2009).
16. Campbell, B. E. *et al.* Norcantharidin analogues with nematocidal activity in *Haemonchus contortus*. *Bioorg. Med. Chem. Lett.* **21,** 3277–3281 (2011).
17. Campbell, B. E., Hofmann, A., McCluskey, A. & Gasser, R. B. Serine/threonine phosphatases in socioeconomically important parasitic nematodes—prospects as novel drug targets? *Biotechnol. Adv.* **29,** 28–39 (2011).
18. Renslo, A. R. & McKerrow, J. H. Drug discovery and development for neglected parasitic diseases. *Nature Chem. Biol.* **2,** 701–710 (2006).
19. Kaminsky, R. *et al.* A new class of anthelmintics effective against drug-resistant nematodes. *Nature* **452,** 176–180 (2008).
20. Maizels, R. M. & Yazdanbakhsh, M. Immune regulation by helminth parasites: cellular and molecular mechanisms. *Nature Rev. Immunol.* **3,** 733–744 (2003).
21. Geenen, P. L. *et al.* The morphogenesis of *Ascaris suum* to the infective third-stage larvae within the egg. *J. Parasitol.* **85,** 616–622 (1999).
22. Bargmann, C. I. Chemosensation in *C. elegans* in *Wormbook* (ed. The *C. elegans* Research Community) (2006); http://www.wormbook.org.
23. Keiser, J. & Utzinger, J. The drugs we have and the drugs we need against major helminth infections. *Adv. Parasitol.* **73,** 197–230 (2010).
24. Harder, A. *et al.* Cyclooctadepsipeptides—an anthelmintically active class of compounds exhibiting a novel mode of action. *Int. J. Antimicrob. Agents* **22,** 318–331 (2003).
25. Heasman, J. Morpholino oligos: making sense of antisense? *Dev. Biol.* **243,** 209–214 (2002).
26. Geldhof, P. *et al.* RNA interference in parasitic helminths: current situation, potential pitfalls and future prospects. *Parasitology* **134,** 609–619 (2007).
27. Lee, I. *et al.* A single gene network accurately predicts phenotypic effects of gene perturbation in *Caenorhabditis elegans*. *Nature Genet.* **40,** 181–188 (2008).
28. Campbell, W. C., Fisher, M. H., Stapley, E. O., Albers-Schonberg, G. & Jacob, T. A. Ivermectin: a potent new antiparasitic agent. *Science* **221,** 823–828 (1983).
29. Berriman, M. *et al.* The genome of the blood fluke *Schistosoma mansoni*. *Nature* **460,** 352–358 (2009).

©2011 Macmillan Publishers Limited. All rights reserved

30. Chen, L., Wilson, D. J., Labello, N. P., Jayaram, H. N. & Pankiewicz, K. W. Mycophenolic acid analogs with a modified metabolic profile. *Bioorg. Med. Chem.* **16,** 9340–9345 (2008).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This project was funded by the Australian Research Council. This research was supported by a Victorian Life Sciences Computation Initiative (grant number VR0007) on its Peak Computing Facility at the University of Melbourne, an initiative of the Victorian Government. Other support from the Australian Academy of Science, the Australian-American Fulbright Commission, Melbourne Water Corporation, and the IBM Research Collaboratory for Life Sciences—Melbourne is gratefully acknowledged. P.W.S. is an investigator with the Howard Hughes Medical Institute. A.R.J. held a CDA1 (Industry) from the National Health and Medical Research Council of Australia. We are indebted to the faculty and staff of the BGI-Shenzhen, who contributed to this study. We also acknowledge the contributions of staff at WormBase (www.wormbase.org).

Author Contributions R.B.G., N.D.Y., B.E.C., J.V., T.W. and P.G. provided the samples and purified nucleic acids for sequencing, X.X. performed the whole genomic amplification of genomic DNA for the large insert libraries. S.L., L.Y., N.Z., A.R.J., Z.X., R.S.H., Y.K. and F.C. undertook the sequencing, assembly and annotation of genomic and transcriptomic data. A.R.J., B.L., Z.X., N.D.Y., Y.L., R.S.H., E.M.S., G.Z., X.F., S.L., F.C. and C.C. planned and performed additional bioinformatic analyses. A.R.J., B.L., N.D.Y. and G.A.A. assisted with statistical analyses, A.R.J., P.N., E.M.S., P.W.S. and R.B.G. drafted and edited the manuscript, tables, figures and Supplementary Information. A.R.J., N.D.Y., S.R., J.W. and R.B.G. conceived and planned the project. A.R.J., B.L., N.D.Y., G.Z., X.F., X.W., J.W., Y.L., H.Y., J.W. and R.B.G. supervised and/or coordinated the research. T.W.H. curated the browsable genome.

Author Information All of the genomic sequence data have been released for public access at WormBase (www.wormbase.org) and are accessible via ftp://ftp.wormbase.org/pub/wormbase/species/a_suum. A browsable genome is also available (via http://www.wormbase.org/db/gb2/gbrowse/a_suum/). Reprints and permissions information is available at www.nature.com/reprints. This paper is distributed under the terms of the Creative Commons Attribution-Non-Commercial-Share Alike licence, and is freely available to all readers at www.nature.com/nature. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to A.R.J. (ajex@unimelb.edu.au), J.W. (wangjun@genomics.org.cn), J.W. (wangjian@genomics.org.cn) or R.B.G. (robinbg@unimelb.edu.au).

©2011 Macmillan Publishers Limited. All rights reserved

METHODS

Sample procurement, preparation and storage. All specimens of *A. suum* were collected from pigs (*Sus scofra*) with naturally acquired infections in Victoria, Australia (adult nematodes) and Ghent, Belgium (larval stages). L3s and L4s were also collected from the liver or lung and from the small intestine, respectively, of pigs, using established procedures[31,32]. Nematodes were washed extensively in sterile physiological saline (37 °C), snap-frozen in liquid nitrogen and then stored at −70 °C until use.

DNA isolation, sequencing and quality control. Total genomic DNA was isolated from the reproductive tract of a single adult female of *A. suum* using a sodium-dodecyl sulphate/proteinase K digestion[33] followed by phenol-chloroform extraction and ethanol precipitation[34]. Total DNA yield was determined using the Qubit fluorometer double-stranded DNA HS Kit (Invitrogen). DNA integrity was verified with a 2100 Bioanalyser (Agilent). Short-insert (170 bp and 500 bp) and mate-pair (800 bp, 2 kb, 5 kb and 10 kb) genomic DNA libraries were prepared and paired-end sequenced using TruSeq chemistry on a HiSeq 2000 (Illumina). Whole-genome amplification, employing the REPLI-g Midi Kit (Qiagen), was used to produce (from 200 ng of genomic template) the required amount of DNA for the construction of the 2-kb, 5-kb and 10-kb libraries (Supplementary Fig. 6). The sequence data generated from each of the six libraries were verified, and low-quality sequences, base-calling duplicates and adapters removed. The size of the genome and the heterozygosity rate were estimated by establishing the frequency of occurrence of each 17-bp *k*-mer (a unique sequence of *k* (that is, 17) nucleotides in length) within the genomic sequence data set (from the 170-bp library) using an established method[5]. Genome size was estimated using a modification of the Lander–Waterman algorithm[35], where the haploid genome length in base pairs is $G = (N \times (L - K + 1) - B)/D$, where N is the read length sequenced in base pairs, L is the mean length of sequence reads, K is the *k*-mer length (17 bp) and B is the number of *k*-mers occurring less than four times (Supplementary Fig. 7). Heterozygosity was evaluated throughout the genome assembly by assessing the distribution of the *k*-mer frequency in the sequence data set.

RNA isolation, sequencing and assembly. We obtained total RNAs from egg-L3s ($n \approx 500{,}000$), liver-L3s ($n \approx 60{,}000$), lung-L3s ($n \approx 80{,}000$) or L4s ($n \approx 30{,}000$) and from the somatic musculature or reproductive tract of each of two adult male and two adult female *A. suum* using the TriPure reagent (Roche), and both yield and quality were verified by 2100 BioAnalyser (Agilent). Polyadenylated (polyA+) RNA was purified from 10 μg of total RNA using Sera-mag oligo(dT) beads, fragmented to a size of 300–500 bp, reverse-transcribed using random hexamers, end-repaired and adaptor-ligated, according to the manufacturer's protocol (Illumina). Ligated products of ~400 bp were excised from agarose and then PCR-amplified (15 cycles), as recommended. Products were purified over a MinElute column (Qiagen) and subjected to paired-end RNA-seq using TruSeq chemistry on a HiSeq 2000 (Illumina) and assessed for quality and adaptor sequence. Transcripts were assembled from RNA-seq data using Oases[36]. All transcripts were used to assess the completeness of the genome assembly and to predict genes.

Genomic assembly and quality control. Following sequencing, all DNA-sequence reads were corrected based on *k*-mer (=17) distribution[5]. Briefly, sequence reads were removed if >10% of bases were ambiguous (represented by the letter N) or multiple adenosine monophosphates (poly-A), and all remaining reads were filtered on the basis of Phred quality. For small insert-size libraries (that is, <800 bp), additional reads were removed from the final data set if >65% of bases were of a low Phred quality (<8). For large insert libraries (2 kb, 5 kb and 10 kb), reads were removed from the final data set if >80% of bases were of a low Phred quality (<8). Duplicate (that is, identical) reads and partial reads representing the Illumina adaptor sequence were also removed, as were reads from the 500-bp library representing paired reads found to overlap by >10 bp (allowing for a 10% mismatch). Corrected and filtered data were assembled into contigs using SOAPdenovo[5], and joined iteratively into scaffolds using a step-wise process (see Supplementary Fig. 8), using the paired reads generated from each library; local assemblies were used to close all gaps. Each nucleotide position in the final assembly was assessed for accuracy by aligning all filtered reads to the scaffolds using SOAP2aligner[37], allowing for up to five mismatches per read. The depth of coverage and repeat content were assessed initially by sliding-window analysis and presented as a frequency distribution (Supplementary Fig. 9). GC-content was estimated using 10-kb non-overlapping sliding windows, and GC-bias[38] was assessed based on a frequency distribution of these data (Supplementary Fig. 10). To assess the completeness of the genome assembly, RNA-seq data representing each of the organs (that is, musculature and reproductive tract), genders and/or stages of *A. suum* sequenced were mapped to the final assembly using the BLAST-like Alignment Tool (BLAT)[39].

Assessment of repeat content and annotation of non-coding RNA. Following genome assembly, tandem repeats were identified using the Tandem Repeats Finder program[40]. Transposable elements were predicted using a combination of homology-based comparisons (using RepeatMasker[41]) and *de novo* approaches (using LTR_FINDER[42], PILER[43] and RepeatScout[44]), with a consensus population of predicted repetitive elements, constructed in RepeatScout using fit-preferred alignment scores. Low-frequency repeats (≤25) and multi-copy genes (in the repeat element library) were filtered using RepeatMasker, producing a non-redundant sequence file, which was then used to identify and classify additional homologous repeats in the genome.

Gene prediction, and synteny and genetic variation analysis. The *A. suum* protein-coding gene set was inferred using *de novo*-, homology- and evidence-based (that is, transcriptomic) approaches. *De novo* gene prediction was performed on a repeat-masked genome using three programs (Augustus, GlimmerHMM and SNAP)[5]; training models were generated from a subset of the transcriptomic data set representing 1,355 distinct genes. Homology-based prediction was conducted by comparison with complete genomic data for *Caenorhabditis elegans*[3], *Pristionchus pacificus*[8] and *Brugia malayi*[9] using a multi-phase strategy, in which (1) all putative homologous gene sequences were preliminarily identified from alignments with protein sequences representing the complete gene set of each of the reference genomes (the longest transcripts were chosen to represent each gene) by TblastN (*e*-value cut-off: 10^{-5}) and grouped into gene-like structures using genBlastA[45]; (2) regions representing these putative genes, and flanking regions (3,000 bp) at the 5′- and 3′-ends of each predicted gene, were extracted from the assembly and aligned to the 'parent' sequences derived from the reference genomes using Genewise[46]; (3) all single-exon genes predicted to have arisen from a retro-transposition and containing at least one frame-shift error or representing incomplete coding domains of <150 bp as well as all multi-exon genes containing more than two frame-shift errors and/or representing incomplete coding domains of <100 bp, were discarded. Evidence-based gene prediction was conducted by aligning all RNA-seq data generated herein against the assembled genome using TopHat[47], with cDNAs predicted from the resultant data using Cufflinks[48]. Following the prediction of genes, a non-redundant gene set representing homology-based, *de-novo*-predicted and RNA-seq-supported genes, was generated using Glean (http://sourceforge.net/projects/glean-gene)[5]. All Glean-predicted genes were retained, as were all genes supported by RNA-seq data and those predicted using two or more *de novo* methods (that is, Augustus, GlimmerHMM and/or SNAP). The open reading frame of each gene was predicted using BestORF (www.softberry.com). To assess the quality and accuracy of the predicted gene set, we examined the length-distribution of all genes, coding sequences, exons and introns, and the distribution of exon numbers for individual genes, and then compared these parameters with those calculated for the published gene sets of *B. malayi*, *C. elegans*, *P. pristionchus* and *M. incognita* (Supplementary Fig. 4).

Following prediction of the finalized gene set, we conducted pairwise analysis of the overall synteny existing between/among the large (>1 Mb) assembly scaffolds for *B. malayi* and *A. suum* relative to the complete *C. elegans* chromosomes. This analysis was undertaken by conducting pairwise alignments among all *A. suum* or *B. malayi* (WS220 assembly: ftp://ftp.sanger.ac.uk/pub2/wormbase/releases/WS220/genomes/b_malayi/) scaffolds larger than 1 Mb in size and the *C. elegans* chromosomes using LASTz (http://www.bx.psu.edu/miller_lab/dist/README.lastz-1.02.00/README.lastz-1.02.00a.html), which were then joined using CHAINNET[49] and output as a .axt alignment from which large-syntenic regions were defined. The resulting alignment files were used to construct synteny images on scaleable vector graphics format using customized perl scripts (ZX). In addition, gene-level synteny analyses were conducted for one-to-one orthologous genes colocalizing to large *A. suum* or *B. malayi* assembly scaffolds (>1 Mb) according to ref. 9. Orthology was determined by pairwise reciprocal BLASTx comparisons between *A. suum*, *B. malayi* and *C. elegans* according to ref. 11. One-to-one orthologous genes shared between either *A. suum* or *B. malayi* and *C. elegans* but not shared among *A. suum* and *B. malayi* based on reciprocal BLASTp analysis were further confirmed by Hidden Markov Modelling using the *jackhmmr* command in the program HMMER 3.0 (ref. 50) and a highly permissive threshold (HMM cutoff: 10^{-2}).

We assessed the genome-wide variation in the exonic regions by mapping all raw reads from our transcriptomic data to the genomic coding domains using Maq[51], and calling SNPs with a minimum coverage threshold of ten reads. All mapped reads were assessed as synonymous (non-coding change), non-synonymous (coding change) or ambiguous (a SNP that was represented in our data set as an ambiguous IUPAC code wherein one nucleotide change would cause a synonymous mutation and the other a non-synonymous mutation) using a custom Perl script (snp_analysis.pl). All genes were then ranked based on their accumulation of SNPs to assess and identify their levels of conservation/variation relative to their

©2011 Macmillan Publishers Limited. All rights reserved

function. We reasoned that, in addition to the real effects of the variability of each gene on their accumulation of SNPs, these data would be influenced also by the coverage achieved for each gene, which is affected by the number of reads available for each gene (that is, their relative levels of transcription) and the length of each gene. Thus, before ranking, the SNP data for each gene was normalized for its calculated reads per kilobase per million reads (RPKM) and total gene length using the simple equation: SNPs per read per kilobase = total SNPs divided by RPKM divided by gene length (in bp) multiplied by 1,000 bp. Following ranking, we explored function among the 2.5% most variable (with the highest rankings based on normalized SNP data) and most conserved genes (with the lowest rankings based on normalized SNP data). Noting the potential inaccuracy associated with estimating the normalized SNP rankings of lowly transcribed genes (owing to a lack of data/coverage), only genes for which at least 100 reads were available were considered in these functional comparisons.

Functional annotation of coding genes. Following the prediction of the protein-coding gene set, each inferred amino acid sequence was assessed for conserved protein domains in the SProt, Pfam, PRINTS, PROSITE, ProDom and SMART databases using InterProScan[52], employing default settings. Gene ontology categories[53] were assigned to each contig inferred to contain at least one conserved protein domain. Gene ontology categories were summarized and standardized to level 2 and level 3 terms, defined using the GOslim hierarchy[54] using WEGO[55]. To characterize further the contigs/transcripts from *A. suum*, we conducted a series of high-stringency BLASTp homology searches (*e*-value cut-off: 10^{-5}) against a variety of databases. Each contig was assessed for a known functional orthologue, defined using the Kyoto Encyclopaedia of Genes and Genomes (KEGG) (www.kegg.com). Where appropriate, orthologous matches were mapped visually to a defined pathway using the KEGG pathway tool (available via www.kegg.com) or clustered to a known protein family using the KEGG-BRITE hierarchy tool (available via www.kegg.com). In addition, the amino acid sequence inferred from each *A. suum* coding gene was compared by BLASTp with protein sequences available for key nematode species (*B. malayi*, *C. elegans*, *P. pacificus* and *M. incognita*) as well as for *Drosophila melanogaster*[4] and *Mus musculus*[56] and those contained within the UniProt[57], SwissProt and TREMBL databases[58]. Key protein groups (for example, peptidases, kinases, phosphatases, GTPases, GPCRs, and transport and channel proteins) were characterized by high-stringency BLASTp homology searching (*e*-value cut-off $<10^{-5}$) of manually curated information sequence data available in the MEROPS[59], WormBase, KS-Sarfari (https://www.ebi.ac.uk/chembl/sarfari/kinasesarfari) and GPCR-Sarfari (https://www.ebi.ac.uk/chembl/sarfari/gpcrsarfari) and the Transporter Classification database[60]. E/S proteins were predicted using Phobius[61], employing both the neural network and hidden Markov models, and by BLASTp homology-searching of the validated signal peptide database[62] and an E/S database containing published proteomic data for *B. malayi*[63,64], *Schistosoma mansoni*[65] and *M. incognita*[66]. In the final annotation, proteins inferred from genes were classified based on a homology match (*e*-value cut-off: $\leq 10^{-5}$) to: (1) a curated, specialist protein database, followed by (2) the KEGG database, followed by (3) the UniProt/SwissProt/TREMBL databases, followed by (4) the annotated gene set for a model organism, including *C. elegans*, *D. melanogaster*, *M. musculus* or *S. cerevisiae*, followed by (5) the gene ontology classification, and, finally, (6) a recognized, conserved protein domain based on InterProScan analysis. Any inferred proteins lacking a match (BLASTp cut-off $\leq 10^{-5}$) in at least one of these analyses were designated hypothetical proteins. The final annotated protein-coding gene set for *A. suum* is available for download at WormBase (in nucleotide and amino acid formats).

Differential transcription analysis. Following RNA-seq, all paired-end reads for each library constructed were aligned to the predicted *A. suum* gene set using TopHat, and quantitative levels of transcription (RPKM)[67] were calculated using Cufflinks. Differential transcription was assessed[68] using a *P*-value cut-off of ≤ 0.01 and a minimum, two-fold difference in absolute RPKM values. False discovery rates for differential transcription were determined[68]. To allow the rapid visual assessment of the statistically significant changes in transcription of each gene between and among individual libraries, we constructed heat-maps representing absolute differences in the RPKM values, calculated for each transcript using a customized Perl script (express_heatmap_RPKM.pl). Genetic interaction networks were predicted[69] based on data available for homologous genes in *C. elegans* (inferred from BLASTp comparisons) and viewed using the program BioLayout 3D[70].

Essentiality and druggability predictions. *A. suum* genes with homology to those in the *C. elegans* and/or *D. melanogaster* genomes were inferred based on BLASTp comparisons using the predicted protein sequences for individual species (*e*-value cut-off 10^{-5}). Phenotypic data for each *C. elegans* and *D. melanogaster* homologue were sourced from WormBase and FlyBase (www.flybase.org), respectively. *A. suum* genes determined[71] to have homologues with lethal phenotypes in both *C. elegans* and *D. melanogaster* were inferred to represent essential genes. Metabolic chokepoints were defined[29,72] and assessed based on *A. suum* gene sequences determined, by BLASTp comparison (10^{-5}), to have an orthologue in the KEGG database. All 'essential' homologues and/or molecules in 'chokepoints' were then queried against the BRENDA[73] and CHEMBL databases (accessible via https://www.ebi.ac.uk/chembldb/), to identify known chemical inhibitors.

Additional bioinformatic analyses, and use of software. Data analysis was conducted in a Unix environment or Microsoft Excel 2007 using standard commands. Bioinformatic scripts required to facilitate data analysis were designed using Perl, BioPerl, Java and Python and are available via http://research.vet.unimelb.edu.au/gasserlab/.

31. Cantacessi, C. *et al.* Differences in transcription between free-living and CO_2-activated third-stage larvae of *Haemonchus contortus. BMC Genom.* **11,** 266, doi:10.1186/1471-2164-11-266 (2010).
32. Saeed, I., Roepstorff, A., Rasmussen, T., Hog, M. & Jungersen, G. Optimization of the agar-gel method for isolation of migrating *Ascaris suum* larvae from the liver and lungs of pigs. *Acta Vet. Scand.* **42,** 279–286 (2001).
33. Gasser, R. B. *et al.* Single-strand conformation polymorphism (SSCP) for the analysis of genetic variation. *Nature Protocols* **1,** 3121–3128 (2007).
34. Sambrook, J. & Russell, D. W. *Molecular Cloning: A Laboratory Manual* 3rd edn, Vol. 3, E.3–E.4 (Cold Spring Harbor Laboratory, 2001).
35. Lander, E. S. & Waterman, M. S. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* **2,** 231–239 (1988).
36. Zerbino, D. R. & Birney, E. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res.* **18,** 821–829 (2008).
37. Li, R. *et al.* SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25,** 1966–1967 (2009).
38. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456,** 53–59 (2008).
39. Kent, W. J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12,** 656–664 (2002).
40. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27,** 573–580 (1999).
41. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protocols Bioinformatics* Ch. 4.10 (2009).
42. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35,** W265–W268 (2007).
43. Edgar, R. C. & Myers, E. W. PILER: identification and classification of genomic repeats. *Bioinformatics* **21** (Suppl. 1), i152–i158 (2005).
44. Price, A. L., Jones, N. C. & Pevzner, P. A. *De novo* identification of repeat families in large genomes. *Bioinformatics* **21** (Suppl. 1), i351–i358 (2005).
45. She, R., Chu, J. S., Wang, K., Pei, J. & Chen, N. GenBlastA: enabling BLAST to identify homologous gene sequences. *Genome Res.* **19,** 143–149 (2009).
46. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res.* **14,** 988–995 (2004).
47. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25,** 1105–1111 (2009).
48. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnol.* **28,** 511–515 (2010).
49. Kent, W. J., Baertsch, R., Hinrichs, A., Miller, W. & Haussler, D. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl Acad. Sci. USA* **100,** 11484–11489 (2003).
50. Eddy, S. R. A new generation of homology search tools based on probabilistic inference. *Genome Inform.* **23,** 205–211 (2009).
51. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **18,** 1851–1858 (2008).
52. Quevillon, E. *et al.* InterProScan: protein domains identifier. *Nucleic Acids Res.* **33,** W116–W120 (2005).
53. Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* **32,** D258–D261 (2004).
54. Camon, E. *et al.* The Gene Ontology Annotation (GOA) project: implementation of GO in SWISS-PROT, TrEMBL, and InterPro. *Genome Res.* **13,** 662–672 (2003).
55. Ye, J. *et al.* WEGO: a web tool for plotting GO annotations. *Nucleic Acids Res.* **34,** W293–W297 (2006).
56. Chinwalla, A. T. *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420,** 520–562 (2002).
57. Wu, C. H. *et al.* The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res.* **34,** D187–D191 (2006).
58. Boeckmann, B. *et al.* The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31,** 365–370 (2003).
59. Rawlings, N. D., Barrett, A. J. & Bateman, A. MEROPS: the peptidase database. *Nucleic Acids Res.* **38,** D227–D233 (2010).
60. Saier, M. H. Jr, Yen, M. R., Noto, K., Tamang, D. G. & Elkan, C. The Transporter Classification Database: recent advances. *Nucleic Acids Res.* **37,** D274–D278 (2009).
61. Kall, L., Krogh, A. & Sonnhammer, E. L. Advantages of combined transmembrane topology and signal peptide prediction—the Phobius web server. *Nucleic Acids Res.* **35,** W429–W432 (2007).
62. Chen, Y. *et al.* SPD—a web-based secreted protein database. *Nucleic Acids Res.* **33,** D169–D173 (2005).
63. Bennuru, S. *et al. Brugia malayi* excreted/secreted proteins at the host/parasite interface: stage- and gender-specific proteomic profiling. *PLoS Negl. Trop. Dis.* **3,** e410 (2009).

©2011 Macmillan Publishers Limited. All rights reserved

64. Hewitson, J. P. *et al.* The secretome of the filarial parasite, *Brugia malayi*: proteomic profile of adult excretory-secretory products. *Mol. Biochem. Parasitol.* **160,** 8–21 (2008).
65. Cass, C. L. *et al.* Proteomic analysis of *Schistosoma mansoni* egg secretions. *Mol. Biochem. Parasitol.* **155,** 84–93 (2007).
66. Bellafiore, S. *et al.* Direct identification of the *Meloidogyne incognita* secretome reveals proteins with host cell reprogramming potential. *PLoS Pathog.* **4,** e1000192 (2008).
67. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* **5,** 621–628 (2008).
68. Audic, S. & Claverie, J. M. The significance of digital gene expression profiles. *Genome Res.* **7,** 986–995 (1997).
69. Zhong, W. & Sternberg, P. W. Genome-wide prediction of *C. elegans* genetic interactions. *Science* **311,** 1481–1484 (2006).
70. Goldovsky, L., Cases, I., Enright, A. J. & Ouzounis, C. A. BioLayout(Java): versatile network visualisation of structural and functional relationships. *Appl. Bioinform.* **4,** 71–74 (2005).
71. Doyle, M. A., Gasser, R. B., Woodcroft, B. J., Hall, R. S. & Ralph, S. A. Drug target prediction and prioritization: using orthology to predict essentiality in parasite genomes. *BMC Genom.* **11,** 222, doi:10.1186/1471-2164-11-222 (2010).
72. Yeh, I., Hanekamp, T., Tsoka, S., Karp, P. D. & Altman, R. B. Computational analysis of *Plasmodium falciparum* metabolism: organizing genomic information to facilitate drug discovery. *Genome Res.* **14,** 917–924 (2004).
73. Scheer, M. *et al.* BRENDA, the enzyme information system in 2011. *Nucleic Acids Res.* **39,** D670–D676 (2011).

©2011 **Macmillan Publishers Limited. All rights reserved**

OPEN ACCESS J PREV MED HYG 2022; 63(SUPPL. 3): E228-E238

REVIEW

An overview of the genetic aspects of hair loss and its connection with nutrition

NURIYE GOKCE[1], NESLIHAN BASGOZ[1], SERCAN KENANOGLU[1], HILAL AKALIN[1], YUSUF OZKUL[1],
MAHMUT CERKEZ ERGOREN[2,3], TOMMASO BECCARI[4], MATTEO BERTELLI[5,6,7], MUNIS DUNDAR[1,*]
[1] Department of Medical Genetics, Faculty of Medicine, Erciyes University, Kayseri, Turkey; [2] Department of Medical Genetics, Faculty of Medicine, Near East University, Nicosia, Cyprus; [3] DESAM Institute, Near East University, Nicosia, Cyprus; [4] Department of Pharmaceutical Science, University of Perugia, Perugia, Italy; [5] MAGISNAT, Peachtree Corners (GA), USA; [6] MAGI Euregio, Bolzano, Italy; [7] MAGI'S LAB, Rovereto (TN), Italy

Keywords

Androgenetic alopecia • Hair loss • Mediterranean Diet • Nutrition

Summary

Hair loss is a widespread concern in dermatology clinics, affecting both men's and women's quality of life. Hair loss can have many different causes, which are critical to identify in order to provide appropriate treatment. Hair loss can happen due to many variables, such as genetic factors or predisposition, vitamin and mineral deficiencies, skin problems, hair growth disorders, poor diet, hormonal problems, certain internal diseases, drug use, stress and depression, cosmetic factors, childbirth, and the chemotherapy process. Treatment for hair loss varies depending on the type of alopecia, deficiency, or excess of structures such as vitamins and minerals, and also on hair and skin structure.

The Mediterranean diet is characterized by low amounts of saturated fat, animal protein, and high amounts of unsaturated fat, fiber, polyphenols, and antioxidants. The main nutrients found in the Mediterranean Diet are rich in antioxidant, anti-inflammatory components. It also has an important place in hair loss treatment, since recently treatment strategies have included polyphenols and unsaturated oils more and more frequently. The goal of this work was to review published articles examining alopecia and its types, the many micronutrients that affect alopecia, and the role of the Mediterranean diet in alopecia. The literature shows that little is known about hair loss, nutritional factors, and diet, and that the data collected are conflicting. Given these differences, research into the function of diet and nutrition in the treatment of baldness is a dynamic and growing topic.

Introduction

Hair loss or alopecia, a clinical condition that is frequently seen in dermatology clinics, can be caused by many etiological factors and it significantly affects the patients' quality of life [1]. This group of diseases is basically divided in two subgroups: cicatricial alopecias and non-cicatricial alopecias. While cicatricial alopecia may progress with loss of follicles, thus causing irreversible hair loss, this condition is usually reversible in non-cicatricial ones. Many causes are known to have a role in non-cicatricial alopecia, including emotional issues, chronic disorders, dietary inadequacies, trace elements, and vitamin deficiencies [2]. Other factors can be stress, drug use, immune system, endocrine disorders, and genetic and epigenetic changes [3].

A balanced and regular diet is very important for healthy hair: sudden weight loss, low-caloric diets, unbalanced diet, obesity, and excessive intake of vitamin and mineral supplements can cause hair loss. Micronutrients, which are the main elements of the hair follicle cycle, are very important in alopecia, which is why dietary supplements (mostly vitamin and mineral) are among the preferred methods to prevent hair loss. Given the frequency of hair loss in current times and its impact on the patients' social lives, finding effective alopecia treatments impacts a huge portion of the population [2].

The role of diet in the development and treatment of alopecia has recently been a hot topic of research. It has been found that plant-rich diets – such as the Mediterranean Diet (MD), whose main nutrients are rich in antioxidants, anti-inflammatory, and estrogenic components – include chemicals that stimulate hair growth and reduce hair loss. These diets contain phytochemicals that promote hair development by lowering the generation of reactive oxygen species in the dermal papilla cells, causing growth hormones to be secreted [4].

This is why dietary practices characterized by a large intake of anti-inflammatory, antioxidant and estrogenic activities are emphasized as additional treatments for alopecia. This review aims to explain the general background of alopecia, to emphasize the role of nutritional and dietary supplements in the treatment of alopecia, and to consider the effects of herbal treatment methods on alopecia patients.

Alopecia and types of alopecia

Alopecia is the partial or complete loss of hair, which can be caused by a disruption in the hair development

https://doi.org/10.15167/2421-4248/jpmh2022.63.2S3.2765

cycle or by damage to the hair follicles as a result of systemic or local factors such as genetics, hormone imbalances, and infection [5-7].

The hair growth cycle consists of four stages: the anagen phase (the growth phase, which is also the longest, lasting 2-7 years); the catagen phase (the transitional phase, lasting up to 2 weeks, which includes hair follicle involution due to apoptosis); the telogen phase (the resting phase, lasting up to 12 weeks); and the exogenous phase (the release phase of telogen hair) [8].

The fact that alopecia is affected by many factors causes the existence of a wide number of types of alopecia. Androgenic alopecia, alopecia areata, chemotherapy-induced alopecia, anagen effluvium, telogen effluvium, traction alopecia, and trichotillomania are some of the most common kinds. The clinical classification system of Rook and Dawber divides all types of alopecia into two categories: cicatricial alopecia and non-cicatricial alopecia [6, 7].

CICATRICIAL ALOPECIAS

Cicatricial alopecias (CAs) are clinical pathological conditions that describe permanent hair loss caused by the replacement of damaged hair follicles with fibrotic scar tissue. They are classified into two groups, primary cicatricial alopecia (PCA) and secondary cicatricial alopecia (SCA) [9, 10].

The PCA group includes multiple inflammatory diseases with distinct clinical and histopathological features and unknown and irreversible causes, primarily affecting and destroying hair follicles. It is responsible for 7% of all hair loss cases. PCA is subdivided into lymphocytic, neutrophilic, or mixed subtypes [11, 12]. Chronic cutaneous lupus erythematosus, lichen planopilaris, classic Brocq pseudopelade, folliculitis decalvans, and dissection folliculitis are some of the diseases with clinical conditions [13].

The SCA group includes inflammatory and neoplastic conditions and physical traumas that usually affect primarily the dermis and cause secondary follicular destruction [9]. While in PCAs the disease directly affects hair follicles, in SCAs they disappear due to secondary reasons. Factors affecting SCA formation include genodermatoses, permanent alopecia due to developmental defects, physical and chemical injuries, infections, inflammatory dermatoses, drug uses, and neoplastic conditions [14]. The classification of cicatricial alopecias is summarized in Table I.

NON-CICATRICIAL ALOPECIAS

Non-cicatricial alopecia is reversible alopecia, characterized by an altered hair cycle and by the fact that hair follicles are preserved [5].

Non-cicatricial alopecias are divided in local alopecias and diffuse alopecias [17]. Local alopecias include alopecia areata, tinea capitis, and trichotillomania. Diffuse alopecias include alopecia totalis or universalis, telogen or anagen effluvium. The classification of non-cicatricial alopecias is summarized in Table II. The most common alopecia in this group is androgenic alopecia.

Androgenic alopecia, caused by the miniaturization of hair follicles, is the most common type of progressive hair loss, affecting 30-50% of men and approximately 30% of middle-aged women. Alopecia areata, the second most common type of non-cicatricial alopecia, is known to be associated with autoimmune problems [5].

Common causes of hair loss

As previously stated, the hair growth cycle consists of four phases: anagen, catagen, telogen, and exogen. Hair loss differs according to the affected stage. Although there are many different reasons for hair loss, androgen and stress-related causes are usually underlying. Other systemic causes of hair loss include telogen effluvium, nutrition, endocrine imbalances, drugs, infections, special diseases, malignancies, problems with the immune system, environmental factors, age, and genetic factors [3].

ANDROGEN-RELATED ALOPECIA

The most common type of hair loss is androgenetic alopecia, which affects 80% of men and 50% of women. In this kind of alopecia, genetic factors and age-related causes affect the androgen mechanism, which is the one allowing vellus hair to turn into longer, thicker, and darker terminal hair: the defect in this mechanism makes it function differently by stimulating the regression of hair follicles to turn into vellus hairs. Androgenetic alopecia is a dynamic and continuous hair loss problem, in which mast cells and lymphocytes are defined around the miniaturized follicle, rich in stem cells [18-20].

Although androgens play an important role in androgenetic alopecia, genetic predisposition is also important: genetic background plays an important role, with a rate of 0.81 in twin studies, thus showing familial clustering. Having a polygenic feature, its genetic background is complex. However, changes in the androgen receptor (AR) gene and 5-alpha reductase gene were found to be effective in androgenetic alopecia. In addition, triple repeat polymorphisms have been associated with single nucleotide polymorphisms. It has been stated that the AR gene on the X chromosome and the ectodysplasin A2 receptor (EDAR2) gene are related and that polymorphisms in this gene are associated with androgenic alopecia, but its exact role has not been fully defined. Androgenetic alopecia is associated with genetic alterations in the WNT signaling pathway, which regulates dermal papilla cells and androgen metabolism. Expression studies and epigenetic studies are limited, as scalp biopsy is difficult to obtain. Overexpression of prostaglandin synthase (PGDS) and the PGDS product prostaglandin D2 (PGD2) restrict hair development by generating an early catagen phase, according to the limited study [8, 21].

STRESS-INDUCED ALOPECIA

Stress is one of the most common reasons of hair growth disorders and hair loss, since it leads to an increase in the level of cortisol released into the body. It has been proven

Tab. I. Classification of cicatricial alopecias [9, 14-16].

Primary Cicatricial Alopecia (working classification of PCA by North American Hair Research Society)	Secondary Cicatricial Alopecia		
Lymphocytic	Chronic cutaneous lupus erythematosus, Lichen planopilaris, Frontal fibrosing alopecia, Graham Little syndrome, Classic pseudopelade (Brocq), Central centrifugal cicatricial alopecia, Alopecia mucinosa, Keratosis follicularis spinulosa decalvans	Physical or chemical trauma	Burns, Toxic/corrosive substances (e.g. acid or alkali burns), Ischemia/pressure, Traction/ trichotillomania (end stage)
		Ionizing radiation	Radiation
		Infections	Bacterial infections, Viral infections, Fungal infections (tinea capitis, in particular, deep trichophytosis caused by zoophilic pathogens)
Neutrophilic	Folliculitis decalvans, Dissecting cellulitis/folliculitis	Malignant and benign tumors	Primary tumors, Metastases, Lymphoproliferative diseases, Epidermal and organoid nevi
		Genodermatoses	Aplasia cutis congenita, Ectodermal dysplasia, Ichthyosis, Epidermolysis bullosa, Darier's disease, Incontinentia pigmenti, Hyalinosis cutis et mucosae
		Granulomatous diseases	Sarcoidosis, Necrobiosis lipoidica
Mixed	Folliculitis (acne) keloidalis, Folliculitis (acne) necrotica, Erosive pustular dermatosis	Autoimmune diseases	Graft–versus–host disease, Scleroderma (en coup de sabre), Lichen sclerosus, Blistering dermatoses (cicatricial pemphigoid)
Non-Specific	Idiopathic scarring alopecia with inconclusive clinical and histopathological findings, End stage of various inflammatory scarring alopecias	"Deposition" dermatoses	Amyloidosis, Mucinosis
		Inflammatory diseases	Psoriasis, Pityriasis amiantacea

that cortisol has a negative effect on the formation mechanism of the hair follicle by breaking down hyaluronan and proteoglycans, which are integrating substances in the extracellular matrix and skin. Acute and chronic stress is known as the main cause of telogen effluvium. Stress also can aggravate the types of alopecia that are primarily caused by endocrine imbalances, immunological responses, and toxic causes. In addition, the stress that occurs in response to hair loss causes hair loss to continue. In animal studies, chronic stress has been associated with hair growth arrest, increased granulation of mast cells, and perifollicular inflammation. Further studies have also shown that certain stress mediators, such as substance P, adrenocorticotropic hormone, prolactin, and cortisol, inhibit hair growth [22].

TELOGEN EFFLUVIUM

Telogen effluvium is a scarless hair loss condition caused by physiological stress (such as delivery, rapid weight loss, mental stress, long-term drug usage), medical conditions like hypo/hyperthyroidism, post-diet effects, and prolonged fasting. Although it is generally reversible, it affects the psychosocial status of the patients, and its mechanism has not been discovered yet. Telogen effluvium is difficult to diagnose and treat because there are many factors at play in the etiology of the disease [23, 24].

ENDOCRINE IMBALANCES

The formation of hair follicles occurs as a result of a molecular genetic process regulated before birth. Many hormones and their receptors play important roles in the healthy progression of this process: thyroid hormones, glucocorticoids, insulin-like growth factor-I, and prolactin are all hormones that influence hair development, although androgens are the most important ones. In addition, vitamin D receptors and retinoid X receptors are also effective in postpartum hair growth [21, 25].

IMMUNE SYSTEM

Alopecia areata and primary cicatricial alopecia are two common forms of immune-mediated alopecia. Alopecia areata is a T-cell-mediated autoimmune illness that affects hair follicles in the anagen stage. The JAK-STAT (Janus kinase-signal converter and transcription activator) signaling pathway is involved in its development [26].

The concept of immune privilege is the phenomenon of protection of organs or tissues that can be seen in the corneal tissue, testis, placenta, liver, intestine, brain, hair follicle, and mucosal tissues from autoimmune threats developed by the host's immune system [27].

Hair follicles in the anagen stage are protected from unwanted immune responses by exhibiting immune privilege from the protrusion level to the bulb. The key mechanisms are the lack or reduction of MHC class I and MHC class II expression, signaling that there is no risk with the CD200 signal, and the absence of antigen presentation. In the development of alopecia areata, collapse is observed in the anagen hair bulb. While NKG2D+ T cells and natural killer (NK) cells rapidly increase in the medium, perifollicular mast cells exhibit proinflammatory properties and interact with CD8+ T cells that recognize autoantigens [26-28].

The intense presence of CD8+CD3+ cytotoxic T cells in anagen stage hair follicles and the more targeting of pigmented hair suggest that the targets of cytotoxic T cells are melanocytes and melanogenesis-related autoantigens. The fact that over 100 single nucleotide polymorphisms are associated in studies indicates the importance of the genetic background [26, 28].

Studies have shown that in primary cicatricial alopecia there is a loss of immune privilege in the protrusion region. Although it is not exactly known what causes the loss of immune privilege, it is thought that among the causes might be perifollicular/intrafollicular increase in skin microtrauma, infectious or psycho-emotional stressors, bacterial superantigens, mast cell degranulation, and ectopic interferon γ (IFN-γ) secretion [26].

OTHER CAUSES OF ALOPECIA

Drug-induced alopecia is a type of non-scarring alopecia involving the scalp, which occurs when one of the two main mechanisms of the normal hair growth cycle, anagen effluvium or telogen effluvium, is stopped. When the drug is stopped, the hair loss improves. Other body hairs are rarely affected by drugs [29]. Anti-hypertensives, anti-arrhythmics, statins, anti-metabolites, psychotropic agents, anti-convulsants, anti-coagulants, antiretrovirals, and H2 blockers are among the drugs that cause hair loss [3].

Hair loss is also known to be caused by agents such as bacterial infection Treponema pallidum, viral infection Epstein-Barr virus, HIV, hepatitis C, varicella-zoster virus, and cytomegalovirus [3].

Hair loss and subsequent alopecia are seen as a result of diseases such as liver and kidney failure, rheumatological disorders, discoid lupus erythematosus, connective tissue disorders, amyloidosis, cutaneous sarcoidosis, systemic sarcoidosis [3]. Alopecia can also be seen in systemic and cutaneous hematological malignancies. Although it is usually associated with mycosis fungoides, hair loss is common in cutaneous T-Cell lymphoma [3, 30].

Nutrition and hair loss

The integrity of the normal function of healthy skin and hair depends largely on an adequate and balanced diet. Whether a nutritional imbalance is a general or specific deficiency, an excessiveness of one component over another can compromise the organism's balance. When looking at the studies published in the literature, it is apparent that the nutritional elements that cause hair loss in healthy adults are still to be thoroughly determined.

In the first published reports in this domain, that are animal studies carried out in the first half of the XX century to explain the relationship between nutritional factors and hair loss, Cunningham identified iron deficiency in rats in 1932, Sullivan and Nicholls found riboflavin deficiency in 1941, and Day studies reported that zinc deficiency causes hair loss in both rats and mice [31-35]. The discovery and identification of vitamins related to dietary deficits have had a profound impact on the evolution of nutrient medicine since the first clinical trials. People frequently look for information on vitamin and mineral supplements and diets to prevent or treat many skin disorders, including hair loss. Hair loss greatly affects a person's appearance and personality in physical and social aspects. As a result, hair loss has a negative impact on the standard of living of men and women alike, therefore making the treatment of alopecia essential [36, 37]. Among the about 100,000 hair follicles on the human scalp, 90% of them are in the anagen phase without alopecia, requiring critical nutrients including proteins, vitamins, and minerals to create healthy hair [38]. As a result, trace elements, such as vitamins, are crucial dietary components [39]. Vitamins and trace elements are micronutrients that are critical elements of the diet, even though they are only required in minute amounts. The normal cycle of hair follicles is dependent on micronutrients as they play a role in cell renewal, which is common in rapidly dividing hair follicles [40]. Furthermore, several micronutrients are suggested to minimize oxidative stress because they play a role in alopecia pathogenesis [41]. As a result, a detailed understanding of the significance of these micronutrients could aid research into alopecia prevention and therapy [40].

MICRONUTRIENTS

Vitamin A, which is made up of a set of unsaturated chemical compounds such as retinol, retinal, and retinoic acid, is an essential nutrient that humans cannot generate and thus it must be taken from diet [42]. Vitamin A and its compounds are required for a variety of functions throughout life, including immune function, cell differentiation and proliferation, reproduction, growth, maintenance of epithelial cell integrity, vision, and embryogenesis [43, 44]. Generally, excessive vitamin A consumption can lead to hair loss in most people [45].

In one study, dietary vitamin A activates follicle stem cells by activating the hair cycle developmental and growth phases in a dose-dependent manner (anagen) [46, 47]. For healthy hair, it is important to maintain homeostasis and, accordingly, to maintain an appropriate concentration of the active metabolite [48]. A sufficient level of vitamin A may usually be obtained via a well-balanced diet. One case reported of a 60-year-old man who took too much vitamin A and had non-scarring fronto-central alopecia and a decrease in pubic and axillary hair. These changes were accompanied by drug toxicity associated with excessive intake of vitamin A by the patient [49]. In conclusion, these studies show that there are some optimal levels of vitamin A that are suitable, but very little or too much of this molecule contributes to the formation, maintenance, or advancement of alopecia [40].

The only vitamin B produced by the body is B7 (biotin). The recommended daily amounts of the vitamin B complex, which are pantothenic acid, riboflavin, thiamine, niacin, B6, B12, and folate, can be taken through a balanced diet. Lack of biotin is extremely uncommon in people who eat a healthy, well-balanced diet, and biotin supplements are not necessary [50]. Hair loss has only been linked to B2 (riboflavin), B7 (biotin), B12, and folate deficits. Biotin (vitamin H or B7) is a vitamin B complex with five carboxylases (acetyl-CoA carboxylase, 3-methylcrotonyl-CoA carboxylase, pyruvate carboxylase, and propionyl-CoA carboxylase) [50]. In addition, biotin is effective in cell signaling, histone modifications, and gene regulation [51]. Biotin lack is a condition that can be inherited or acquired; the inheritance pattern is autosomal recessive and there can be neonatal and infantile forms. It is caused by a lack of the enzyme holocarboxylase synthetase and appears from the first six weeks of life; it can be a potentially fatal disease in newborns. Diffuse dermatitis and alopecia are common among survivors. The scalp may be devoid of vellus and terminal follicles, and may also lack lanugo hairs, eyebrows, and eyelashes. Hair on the scalp, eyebrows, and eyelashes may be present but sparse or completely lost [52, 53].

Despite the popularity of biotin, there is still insufficient evidence in randomized controlled trials to support that supplementing this micronutrient prevents or treats hair loss. There are no clinical studies showing that biotin supplements are effective in treating hair loss unless people are deficient [54]. Hair loss, rashes on the skin, and fragile nails are all signs of biotin deficiency. In one study, 38% of women suffering from hair loss had low levels of biotin [55].

Humans cannot produce folate. Folate is essential for cell division and maintenance, and for synthesis of the nucleotide (thymidine), which is required for DNA repair. Folate also plays an important role in the 'site-specific' methylation of the cytosine base in gene expression. Remethylation of plasma homocysteine to methionine is another action of folate/folic acid. This water-soluble vitamin's demand is satisfied in part by dietary folate and in part with the use of synthetic folic acid [56]. Hair, skin, and nail changes may indicate a lack of folate [38]. Vitamin B12 (also known as cobalamin) is needed to activate folate, which is essential for DNA synthesis. Inadequate nutrition causes impaired nerve transmission and insufficient erythrocyte and other hematological cell production. B12 is an essential nutrient in the transfer of a methyl group in a reaction that requires methionine synthase, which converts homocysteine to methionine. This reaction activates folate, which can be used in DNA synthesis. Myelin production, and thus the preservation and repair of nerve axons, both require B12. In addition, B12 is required for energy synthesis in mitochondria and erythropoiesis in the bone marrow [57].

Due to the effect of B12 and folate on nucleic acid, it has been highlighted that they could play a role in the proliferation of hair follicles [38]. Additionally, a case-control review of dietary research on alopecia areata found no changes in vitamin B12 values in patients with and without alopecia areata. Although vitamin B12 appears to not affect hair growth and repair, research on the subject is scarce [53].

Ascorbic acid, often known as vitamin C, is a water-soluble vitamin produced as a product of glucose metabolism. Vitamin C has a chelating and reducing effect. This effect helps intestinal absorption and mobilization of iron [58, 59]. Accordingly, vitamin C supplementation is critical for patients suffering from hair loss due to iron deficiency [38]. The enzyme l-gulonolactone oxidase, which is necessary for the production of vitamin C, is evolutionarily deficient in humans, which is why vitamin C must be received from diet [60]. The pathophysiological role of vitamin C in the hair cycle is poorly known [53]. Citrus juices contain large amounts of vitamin C, and patients with iron deficiency are recommended to take oral supplements such as orange juice [61]. While vitamin C insufficiency has been linked to anomalies in body hair [59], there's still no evidence of a link between vitamin C values and hair loss.

Vitamin D is a steroid hormone that is fat-soluble. Vitamin D plays a regulatory role in various pathways, such as calcium absorption, bone mineralization, immune system, and DNA transcription [62, 63]. Vitamin D, whether obtained from diet or through the skin, is inactive and must be activated by enzymes [38]. Vitamin D insufficiency is a typical occurrence [64]. It was observed that vitamin D has an effect on the hair cycle in animal research. By attaching to the nuclear vitamin D receptor (VDR), vitamin D regulates keratinocyte development and differentiation. VDR immunoreactivity

Tab. II. Classification of Non-Cicatricial Alopecias [17].

Diffuse Alopecia	Local Alopecia
Female pattern hair loss	Alopecia areata
Male pattern hair loss	Tinea capitis
Diffuse alopecia areata	Traction alopecia
Alopecia totalis or universalis	Trichotillomania
Telogen effluvium	
Anagen effluvium	

is highest in the anagen phase in murine hair follicle keratinocytes [38, 53]. Vitamin D thus plays a function in alopecia and may be the cause of hair loss in vitamin D-deficient rickets patients [65]. Alopecia, hypocalcemia, rickets, hyperparathyroidism, and osteomalacia develop in mice lacking the vitamin D receptor (VDR). Normalizing mineral ion homeostasis prevents all these abnormalities, except alopecia. Hair regeneration experiments in athymic nude mice demonstrate that the lack of VDR in keratinocytes results in a defect in anagen initiation, similar to the one observed in mice without VDR. Although these studies show that expression of the VDR in keratinocytes is necessary, they do not prove that it is sufficient to maintain the normal hair cycle [66]. Patients with VDR mutations who develop vitamin D resistance also suffer from severe alopecia of the body and scalp [53, 65]. In two patients with inherited vitamin D-resistant rickets and baldness, Forghani et al. discovered novel nonsense mutations in the VDR gene [67, 68]. Daroach et al. compared patients with alopecia areata to healthy controls in a prospective study to correlate VDR expression in alopecia areata patients, finding higher vitamin D deficiency levels in the alopecia group [69]. To better understand the effects of vitamin D in alopecia, more large-scale research is required [68].

Iron is vital for all cellular functions because of its role in electron transport, oxygen delivery, and enzyme activity. It includes a wide range of enzymes, from hydroxylases – regulating cellular metabolism – to demethylases – changing DNA chromatin and consequently gene expression [70]. Iron is vital in rapidly proliferating cells, such as those in the hair follicle matrix, since it is a cofactor for ribonucleotide reductase, the rate-limiting enzyme in DNA synthesis. Some studies have demonstrated that various genes in the hair follicle are regulated by iron [71]. Iron deficiency, which also causes telogen effluvium, is the most frequent dietary deficiency worldwide [72, 73]. The level of serum ferritin, also known as iron-binding protein, is an important correlate of total body iron stores and is utilized in hair loss research as an indicator. Iron deficiency is frequent in women who are experiencing hair loss [31]. A mutation in the *TMPRSS6* gene caused the "mask mouse" to lose body hair and develop iron deficiency anemia, according to a study conducted by Du et al. A mutation in *TMPRSS6* was linked to an inability to downregulate *HAMP* expression in mice, and elevated hepcidin levels were linked to reduced dietary iron absorption and, as a result, iron insufficiency. Iron treatment, interestingly, corrected iron deficiency and stimulated hair growth in these animals [74]. Further research is needed to develop iron supplementation guidelines, therapy markers, and cures for patients suffering from hair loss due to iron deficiency.

Selenium (Se) is primarily involved in thyroid hormone metabolism, redox homeostasis, and enzymatic activity like glutathione peroxidases (GPx) as an important microelement. The antioxidant potential of other selenoproteins is thought to have a significant impact on human health [75]. Hair loss has been observed in selenium-deficient rats [76] and in knock-out mice lacking a selenium cofactor protein [77]. Although these findings suggest a link between selenium deficiency and hair loss, the role of selenium in human hair follicles is still not fully understood. Hair loss can be caused by selenium toxicity [53]. Among people who have been diagnosed with selenium toxicity, 72 percent was affected by hair loss affected, losing 10 percent to 100 percent of their hair [78]. After starting daily selenium therapy in 6 infants with pseudoalbinism alopecia-like findings, alopecia and pseudoalbinism improved when serum selenium levels reverted to normal [79]. Another clinical study showed that there was a significant reduction in hair loss and other gastrointestinal symptoms in ovarian cancer patients who received selenium supplements during chemotherapy as compared to controls. This result shows that selenium has a beneficial effect during chemotherapy [80]. Selenium is thought to play a role in the hair cycle, and healthy people can obtain enough through their diet. Supplementation is not suggested unless there are known deficits, particularly given the risk of hair loss due to toxicity.

Zinc is a trace element that is essential for a variety of cellular functions through proteins, enzymes, and zinc fingers, and also acts as a multitude of transcription factors important in gene regulation [81, 82]. Zinc deficiency can cause telogen effluvium, thin white and brittle hair, as well as many dermatological problems [83]. Superoxide dismutase is one of the zinc-dependent enzymes, which is thought to be one of the contributing factors to alopecia areata [84]. The function of antioxidant processes in alopecia areata is a popular issue in the scientific community. Zinc has also been linked to the Hedgehog signaling pathway, which has been linked to hair follicle morphogenesis [53]. Although zinc homeostasis impairment has been linked to a number of issues, its diabetic effects and function in metabolic syndrome are still being researched [85]. Zinc is a cofactor for enzymes involved in dopamine transport [86], cell membrane stabilization [87], and prostaglandin metabolism. Data relating zinc levels to telogen effluvium and androgenetic alopecia are not homogeneous. In another study comparing 312 people with hair loss with 32 controls, patients with alopecia areata and telogen effluvium showed low zinc levels [88].

In summary, nutritional supplement can damage the hair, when taken without a cause: a surplus dosage of micronutrients, including vitamins A and E, as well as selenium, has been linked to hair loss [83]. Hypervitaminosis A can cause hair loss, and findings on the effects of isotretinoin on hair loss back up this theory. Even though the connection between vitamin D values and telogen effluvium or androgenetic alopecia is still contested, most researchers concur that patients having alopecia and vitamin D insufficiency should take vitamin D supplements. On the other hand, people with iron-deficient alopecia need vitamin C. There is no evidence that vitamin E has a function in androgenetic alopecia or telogen effluvium. Women with hair loss are more likely to be lacking iron. In individuals with iron insufficiency and/

or low ferritin, many researchers deem that iron supplementation is necessary. Data on zinc concentrations in telogen effluvium and androgenetic alopecia are not yet clear and zinc screening is not advisable. It can be also a cause of hair loss, selenium toxicity, and riboflavin deficiency. Nevertheless, there are still no comprehensive studies to make any recommendations about screening for riboflavin or selenium. Hair loss can be caused by a lack of biotin. However, there is no proof that biotin supplementation increases hair growth. Exogenous biotin may also interfere with several laboratory tests, resulting in erroneous negative or positive results. Although some research has been carried out on the link between hair loss and B12 or folate/folic acid, currently there is not enough data on these dietary supplements.

Mediterranean diet and hair loss treatment

The Mediterranean Diet (MD), among the healthiest nutrition models accepted by the whole world, was put forward by Prof. Ancel Keys and his friends from examining the typical diet in Southern Italy and Greece. The MD is characterized by a low intake of saturated fat and animal protein, by a high intake of unsaturated fats, fiber, and antioxidants, and by appropriate intake of omega-6/omega-3 fatty acids. It is generally rich in vegetables, fruits, legumes, nuts, grains, fish, and unsaturated fats, with small amounts of meat and dairy products [89-91]. In the Seven Country Study, the MD has been linked to a lower risk of coronary artery disease. In addition, several studies support its effectiveness in preventing diabetes, obesity, and even various types of cancer [89, 91]. This is thought to be related to an adequate intake of polyphenols, which play a critical role in the MD. Polyphenols have genomic effects that play a part in the inactivation and activation of regulatory genes. They also exert epigenetic effects by modulating the expression of microRNAs, regulating mechanisms such as DNA methylation and histone modification, and chromatin rearrangement [91, 92]. The genetic and epigenetic effects of polyphenols, one of the main components of the MD, on the genome have recently been the focus of attention for researchers.

In addition, polyphenols protect the body against harsh climatic conditions and oxidative stress, which are involved in the defense against ultraviolet radiation and pathogen attacks in plants. Diets rich in these secondary metabolites found in fruits, vegetables, fiber foods, cereals, and beverages protect against the development of chronic and neurodegenerative diseases, diabetes, cancer, cardiovascular diseases, aging, and hypertension in humans [93]. Polyphenols can act as antioxidants, stimulate cell signaling pathways, and reduce inflammation by blocking gene activity. Studies have proven that the MD plays an important role in reducing inflammation [94]. The presence of over 8,000 polyphenols identified so far and the differences between the regions of the MD are considered limiting factors in the study of the effects of this diet. However, the protective effect of polyphenols has been proven by many studies [91, 93]. Plants containing flavonoids are promising in terms of reducing hair loss or stimulating hair growth [95].

The effects of the main components of the foods consumed in the MD are examined in hair loss and in many other diseases. Studies have shown that the polyphenolic compounds in tea significantly increase hair growth. In addition, it has been reported that epigallocatechin-3-gallate, the main component of tea polyphenols, can reduce the risk of androgenetic alopecia by inhibiting 5α-reductase, which increases hair growth in humans [96].

Essential fatty acids (linolenic, linoleic, oleic, myristoleic, palmitoleic, and stearic acids) found in olive oil, which has an important place in the MD, have inhibitory effects on 5α-reductase (5AR) that provide hair regrowth [4, 18, 97]. It has been reported that hair loss and depigmentation are seen in hair and eyebrows in people affected by linoleic acid and alpha-linolenic acid deficiency [83]. Mice with testosterone-induced alopecia showed a significant increase in hair growth after day 16 of topical applications of hydroalcoholic extracts of rosemary, which is a polyphenol. In vitro tests showed strong inhibition of the 5AR enzyme binding to the dihydrotestosterone (DHT) receptor [98, 99]. Although not fully proven, topical applications of coconut oil for hair loss treatment continue to be used [100].

Polyunsaturated fatty acids are thought to be absorbed into mast cell membranes, altering membrane-associated enzymes such phospholipases, which affect mast cell exocytosis. In a monkey study, it was shown that supplementing polyunsaturated fatty acids reduced hair loss in slightly alopecic monkeys; however, the cause of the hair loss was unknown [101].

A plant nanoparticle of safflower oil body loaded with human fibroblast growth factor 10 (hFGF10) in mice with androgenetic alopecia was found to accelerate hair regeneration by targeting hair follicles and reducing inflammation without any toxicity It has also been found to reduce hair follicle inflammation by inhibiting the overproduction of TNF-α, IL-1β, and IL-6 in macrophages and increasing the proliferation of dermal papilla cells. In this respect, it can be seen as a viable treatment option in androgenetic alopecia [102]. The standardized para rubber seed oil has been proven to be a safe and efficient bio-oil to stimulate hair growth or reduce/suppress hair loss [103].

Although there are many products for the treatment of alopecia, their effectiveness is controversial; research thus focuses on more effective pharmaceutical product development. As a result, the bioactivity of oligomeric procyanidin, a dimeric derivative from apples that can stimulate hair epithelial cell proliferation and induce anagen phase in vitro and in humans, is a popular application [104].

In one study, it was observed that resveratrol, an antioxidant, anti-inflammatory, and anti-apoptotic agent, when applied to mice, stimulated the transition of the hair cycle from the telogen phase to the anagen phase, delayed the transition to the catagen phase, and protected the hair follicle from oxidative damage. Resveratrol is considered to be a potential candidate drug for the treatment of alopecia [105]. In addition, a topical lotion containing Redensyl and Sepicontrol A5 polyphenols is used as an alternative treatment for androgenetic alopecia, as it is very safe and provides patient satisfaction [106]. When given to rabbits, *Ficus benghalensis* leaf extracts were found to be effective in increasing hair follicle development by lowering the action of 5-reductase II [107].

As a result, a lack of certain nutrients, such as vitamins, minerals, vital fatty acids, and proteins, can cause hair loss, structural anomalies, and color changes, albeit the exact causes remain unknown [83]. Hair loss is prevalently treated by dermatologists, and it has a significant psychological and emotional impact on patients. Micronutrients like vitamins and minerals have a vital, though not completely understood, role in hair follicle formation and in immune cell activity. A lack of these micronutrients could be a modifiable risk factor for the development, prevention, and treatment of alopecia.

It has been proven by studies that those who follow the MD have less risk of alopecia. As the use of polyphenol-based phytochemicals – frequently used in the MD– increases, the effect of polyphenols in the treatment of alopecia is taken into consideration. In addition, in several studies essential oils were used in the treatment of alopecia, leading to positive results [4, 91].

Considering all the data, more research is needed soon on the role of nutrients in the hair cycle, its association with known alopecia diseases, and effective supplementation regimens. The fact that the incidence of alopecia is low in those fed the MD is promising for the development of the most appropriate preventive and supportive treatments by further researching the active ingredients frequently used in the MD.

Acknowledgements

None.

Conflicts of interest statement

The authors declare no conflict of interest.

Author's contributions

MD, MB, and TB developed the study design and conceptualization of the research methodology. NG, NB, and SK contributed to the manuscript's writing. MD, YO, HA, and MCE contributed to manuscript reviewing and editing processes. All authors have read and approved the final manuscript

References

[1] Mounsey AL, Reed SW. Diagnosing and treating hair loss. Am Fam Physician 2009;80:356-62.

[2] Almohanna HM, Ahmed AA, Tsatalis JP, Tosti A. The role of vitamins and minerals in hair loss: a review. Dermatol Ther (Heidelb) 2019;9:51-70. https://doi.org/10.1007/s13555-018-0278-6

[3] Lin RL, Garibyan L, Kimball AB, Drake LA. Systemic causes of hair loss. Ann Med 2016;48:393-402. https://doi.org/10.1080/07853890.2016.1180426

[4] Pham CT, Romero K, Almohanna HM, Griggs J, Ahmed A, Tosti A. The role of diet as an adjuvant treatment in scarring and nonscarring alopecia. Ski Appendage Disord 2020;6:88-96. https://doi.org/10.1159/000504786

[5] Cardoso CO, Tolentino S, Gratieri T, Cunha-Filho M, Lopez RFV, Gelfuso GM. Topical treatment for scarring and non-scarring alopecia: An overview of the current evidence. Clin Cosmet Investig Dermatol 2021;14:485-99. https://doi.org/10.2147/CCID.S284435

[6] Rambwawasvika H, Dzomba P, Gwatidzo L. Alopecia types, current and future treatment. J Dermat Cosmetol 2021;5:93-99. https://doi.org/10.15406/jdc.2021.05.00190

[7] Cotsarelis G, Millar SE. Towards a molecular understanding of hair loss and its treatment. Trends Mol Med 2001;7:293-301. https://doi.org/10.1016/S1471-4914(01)02027-5

[8] Martinez-Jacobo L, Villarreal-Villarreal CD, Ortiz-López R, Ocampo-Candiani J, Rojas-Martínez A. Genetic and molecular aspects of androgenetic alopecia. Indian J Dermatol Venereol Leprol 2018;84:263-8. https://doi.org/10.4103/ijdvl.IJDVL_262_17

[9] Fanti PA, Baraldi C, Misciali C, Piraccini BM. Cicatricial alopecia. G Ital Dermatologia Venereol 2018;153:230-42. https://doi.org/10.23736/S0392-0488.18.05889-3

[10] Karakuzu A. Primary cicatricial alopecia. Turkderm 2014;48:56-9. https://doi.org/10.4274/turkderm.48.s13

[11] Civaş E, Aykan A, Aksoy B, Eski M. Surgical correction of primary cicatricial alopecia: Experience with nine patients. TURKDERM 2018;52:85-90. https://doi.org/10.4274/turkderm.20270

[12] Filbrandt R, Rufaut N, Jones L, Sinclair R. Primary cicatricial alopecia: Diagnosis and treatment. Cmaj 2013;185:1579-85. https://doi.org/10.1503/cmaj.111570

[13] Villablanca S, Fischer C, García-García SC, Mascaró-Galy JM, Ferrando J. Primary Scarring Alopecia: Clinical-Pathological Review of 72 Cases and Review of the Literature. Skin Appendage Disord 2017;3:132-43. https://doi.org/10.1159/000467395

[14] Finner AM, Otberg N, Shapiro J. Secondary cicatricial and other permanent alopecias. Dermatol Ther 2008;21:279-94. https://doi.org/10.1111/J.1529-8019.2008.00210.X

[15] Olsen EA, Bergfeld WF, Cotsarelis G, Price VH, Shapiro J, Sinclair R, Solomon A, Sperling L, Stenn K, Whiting DA, Bernardo O, Bettencourt M, Bolduc C, Callendar V, Elston D, Hickman J, Ioffreda M, King L, Linzon C, McMichael A, Miller J, Mulinari F, Trancik R. Summary of North American Hair Research Society (NAHRS)-sponsored Workshop on Cicatricial Alopecia, Duke University Medical Center, February 10 and 11, 2001. J Am Acad Dermatol 2003;48:103-10. https://doi.org/10.1067/MJD.2003.68

[16] Kanti V, Röwert-Huber J, Vogt A, Blume-Peytavi U. Cicatricial alopecia. JDDG 2018;16:435-61. https://doi.org/10.1111/DDG.13498

[17] Mounsey AL, Reed SW. Diagnosing and treating hair loss. Am Fam Physician. 2009;80:356-62.

[18] Lourith N, Kanlayavattanakul M. Hair loss and herbs for treatment. J Cosmet Dermatol 2013;12:210-22. https://doi.org/10.1111/JOCD.12051

[19] Gentile P, Garcovich S. Advances in regenerative stem cell therapy in androgenic alopecia and hair loss: wnt pathway, growth-factor, and mesenchymal stem cell signaling impact analysis on cell growth and hair follicle development. Cells 2019;8:466. https://doi.org/10.3390/cells8050466

[20] Alessandrini A, Bruni F, Piraccini BM, Starace M. Common

causes of hair loss- clinical manifestations, trichoscopy and therapy. J Eur Acad Dermatol Venereol 2021;35:629-40. https://doi.org/10.1111/jdv.17079

[21] Batrinos ML. The endocrinology of baldness. Hormones 2014;13:197-212. https://doi.org/10.1007/BF03401334

[22] Thom E. Stress and the hair growth cycle: Cortisol-induced hair growth disruption. J Drugs Dermatology 2016;15:1001-4.

[23] Chien Yin GO, Siong-See JL, Wang ECE. Telogen Effluvium – a review of the science and current obstacles. J Dermatol Sci 2021;101:156-63. https://doi.org/10.1016/J.JDERMSCI.2021.01.007

[24] Asghar F, Shamim N, Farooque U, Sheikh H, Aqeel R. Telogen Effluvium: A Review of the Literature. Cureus 2020;12:e8320. https://doi.org/10.7759/cureus.8320

[25] Alonso LC, Rosenfield RL. Molecular genetic and endocrine mechanisms of hair growth. Horm Res 2003;60:1-13. https://doi.org/10.1159/000070821

[26] Anzai A, Hsi E, Wang C, Lee EY, Aoki V, Christiano AM. Pathomechanisms of immune-mediated alopecia. Int Immunol 2019;31:439-47. https://doi.org/10.1093/intimm/dxz039

[27] Bertolini M, McElwee K, Gilhar A, Bulfone-Paus S, Paus R. Hair follicle immune privilege and its collapse in alopecia areata. Exp Dermatol 2020;29:703-25. https://doi.org/10.1111/EXD.14155

[28] Rahmani W, Sinha S, Biernaskie J. Immune modulation of hair follicle regeneration. npj Regen Med 2020;5. https://doi.org/10.1038/s41536-020-0095-2

[29] Patel M, Harrison S, Sinclair R. Drugs and Hair Loss. Dermatol Clin 2013;31:67-73. https://doi.org/10.1016/j.det.2012.08.002

[30] Bi MY, Curry JL, Christiano AM, Hordinsky MK, Norris DA, Price VH, Duvic M. The spectrum of hair loss in patients with mycosis fungoides and Sézary syndrome. J Am Acad Dermatol 2011;64:53-63. https://doi.org/10.1016/j.jaad.2009.12.056

[31] Rushton DH. Nutritional factors and hair loss. Clin Exp Dermatol 2002;27:396-404. https://doi.org/10.1046/j.1365-2230.2002.01076.x

[32] Cunningham IJ. Influence of dietary iron on hair and wool growth. https://www.cabdirect.org/cabdirect/abstract/19321401596 (accessed on: April 26, 2022).

[33] Sullivan M, Nicholls J. Nutritional dermatosis in the rat. IV. Riboflavin deficiency. J Invest Dermatol 1941;4:181-91. https://doi.org/10.1038/jid.1941.17

[34] Day HG. The effects of zinc deficiency in the mouse. https://www.cabdirect.org/cabdirect/abstract/19421402451 (accessed on: April 26, 2022).

[35] Day HG, McCollum EV. Effects of Acute Dietary Zinc Deficiency in the Rat. Proc Soc Exp Biol Med 1940;45:282-4. https://doi.org/10.3181/00379727-45-11653P

[36] Lohia K, Doshi BR, Manjunathswamy B. Hair loss severity and its impact on quality of life in patients suffering from androgenic alopecia: A one-year cross-sectional study. Clin Dermatology Rev 2022;5:59. https://doi.org/10.4103/CDR.CDR_52_20

[37] Adamowicz R, Załe P, Dukiel A, Nowicka D. Association between androgenetic alopecia and psychosocial disease burden: a cross-sectional survey among polish men. 2022;2022:1845044. https://doi.org/10.1155/2022/1845044

[38] Almohanna HM, Ahmed AA, Tsatalis JP, Tosti A. The role of vitamins and minerals in hair loss: a review. Dermatol Ther (Heidelb) 2019;9:51-70. https://doi.org/10.1007/S13555-018-0278-6

[39] Mason JB. Vitamins, Trace Minerals, and Other Micronutrients. Twenty Fourth edition. Elsevier Inc. 2012. https://doi.org/10.1016/b978-1-4377-1604-7.00558-3

[40] Thompson JM, Mirza MA, Park MK, Qureshi AA, Cho E. The role of micronutrients in alopecia areata: a review. Am J Clin Dermatol 2017;18:663-79. https://doi.org/10.1007/S40257-017-0285-X

[41] Prie BE, Voiculescu VM, Ionescu-Bozdog OB, Petrutescu B, Iosif L, Gaman LE, Clatici VG, Stoian I, Giurcaneanu C. Oxidative stress and alopecia areata. J Med Life 2015;8:43. https://doi.org/10.1016/j.jaad.2009.10.0

[42] Tanumihardjo SA, Russell RM, Stephensen CB, Gannon BM, Craft NE, Haskell MJ, Lietz G, Schulze K, Raiten DJ. Biomarkers of Nutrition for Development (BOND) – Vitamin A Review. J Nutr 2016;146:1816S. https://doi.org/10.3945/JN.115.229708

[43] Duester G. Retinoic acid synthesis and signaling during early organogenesis. Cell 2008;134:921-31. https://doi.org/10.1016/J.CELL.2008.09.002

[44] Stephensen CB. Vitamin A, infection, and immune function. Annu Rev Nutr 2001;21167-92. https://doi.org/10.1146/ANNUREV.NUTR.21.1.167

[45] Yamamoto K, Sadahito K, Yoshikawa M, Nobuyuki O, Mikami O, Yamada M, Nakamura K, Yasuyuki N. Hyena disease (premature physeal closure) in calves due to overdose of vitamins A, D3, E. Vet Hum Toxicol 2003;45:85-7.

[46] Suo L, Sundberg JP, Everts HB. Dietary vitamin A regulates wingless-related MMTV integration site signaling to alter the hair cycle. Exp Biol Med 2015;240:618. https://doi.org/10.1177/1535370214557220

[47] Duncan FJ, Silva KA, Johnson CJ, King BL, Szatkiewicz JP, Kamdar SP, Ong DE, Napoli JL, Wang J, King LE, Whiting DA, McElwee KJ, Sundberg JP, Everts HB. Endogenous retinoids in the pathogenesis of alopecia areata. J Invest Dermatol 2013;133:334. https://doi.org/10.1038/JID.2012.344

[48] Everts, H.B. Endogenous retinoids in the hair follicle and sebaceous gland. Biochim Biophys Acta 2012;1821:222-9. https://doi.org/10.1016/j.bbalip.2011.08.017

[49] Cheruvattath R, Orrego M, Gautam M, Byrne T, Alam S, Voltchenok M, Edwin M, Wilkens J, Williams JW, Vargas HE. Published online in Wiley InterScience (www.interscience.wiley.com). LIVER Transplant 2006;121888-91. https://doi.org/10.1002/lt.21007

[50] Saleem F, Soos MP. Biotin Deficiency. J Nutr Educ 2021;13:96. https://doi.org/10.1016/S0022-3182(81)80040-4

[51] Zempleni J, Kuroishi T. Biotin. Adv Nutr 2012;3:213. https://doi.org/10.3945/AN.111.001305

[52] Goldberg LJ, Lenzy Y. Nutrition and hair. Clin Dermatol 2010;28:412-9. https://doi.org/10.1016/J.CLINDERMATOL.2010.03.038

[53] O'Connor K, Goldberg LJ. Nutrition and hair. Clin Dermatol 2021;39:809-18. https://doi.org/10.1016/J.CLINDERMATOL.2021.05.008

[54] Guo EL, Katta R. Diet and hair loss: effects of nutrient deficiency and supplement use. Dermatol Pract Concept 2017;7:1-10. https://doi.org/10.5826/dpc.0701a01

[55] Trüeb RM. Serum Biotin Levels in Women Complaining of Hair Loss. Int J Trichology 2016;8:73-7. https://doi.org/10.4103/0974-7753.188040

[56] Czeizel AE, Dudás I, Vereczkey A, Bánhidy F. Folate deficiency and folic acid supplementation: the prevention of neural-tube defects and congenital heart defects. Nutrients 2013;5:4760. https://doi.org/10.3390/NU5114760

[57] Pawlak R, Parrott SJ, Raj S, Cullum-Dugan D, Lucus D. How prevalent is vitamin B(12) deficiency among vegetarians? Nutr Rev 2013;71:110-7. https://doi.org/10.1111/NURE.12001

[58] Valdés F. Vitamin C. Actas Dermosifiliogr 2006;97:557-68. https://doi.org/10.1016/S0001-7310(06)73466-4

[59] Fleming JD, Martin B, Card DJ, Mellerio JE. Pain, purpura and curly hairs. Clin Exp Dermatol 2013;38:940-2. https://doi.org/10.1111/CED.12118

[60] Chhabra N, Chhabra S. A case oriented approach towards biochemistry. First edition. New Delhi: Jaypee Brothers Medical Publishers 2013.

[61] Gropper S, Smith J. Advanced human nutrition and metabolism, 6th ed. Belmont: Wadsworth, Cengage Learning 2012.

[62] D'Aurizio F, Villalta D, Metus P, Doretto P, Tozzoli R. Is vitamin D a player or not in the pathophysiology of autoimmune thyroid diseases? Autoimmun Rev 2015;14:363-9. https://doi.org/10.1016/J.AUTREV.2014.10.008

[63] Antico A, Tampoia M, Tozzoli R, Bizzaro N. Can supplementation with vitamin D reduce the risk or modify the course of autoimmune diseases? A systematic review of the literature. Autoimmun Rev 2012;12:127-36. https://doi.org/10.1016/J.AUTREV.2012.07.007

[64] Parva NR, Tadepalli S, Singh P, Qian A, Joshi R, Kandala H, Nookala VK, Cheriyath P. Prevalence of vitamin D deficiency and associated risk factors in the US population (2011-2012). Cureus 2018;10:e2741. https://doi.org/10.7759/CUREUS.2741

[65] Vupperla D, Lunge S, Elaprolu P. Vitamin D-dependent rickets type ii with alopecia: a rare case report. Indian J Dermatol 2018;63:176-9. https://doi.org/ 10.4103/IJD.IJD_434_17

[66] Chen CH, Sakai Y, Demay MB. Targeting expression of the human vitamin D receptor to the keratinocytes of vitamin D receptor null mice prevents alopecia. Endocrinology 2001;142:5386-9. https://doi.org/10.1210/ENDO.142.12.8650

[67] Forghani N, Lum C, Krishnan S, Wang J, Wilson DM, Blackett PR, Malloy PJ, Feldman D. Two new unrelated cases of hereditary 1,25-dihydroxyvitamin D-resistant rickets with alopecia resulting from the same novel nonsense mutation in the vitamin D receptor gene. J Pediatr Endocrinol Metab 2010;23:843-50. https://doi.org/10.1515/JPEM.2010.136

[68] Gerkowicz A, Chyl-Surdacka K, Krasowska D, Chodorowska G. The role of vitamin D in non-scarring alopecia. Int J Mol Sci 2017;18:2653. https://doi.org/10.3390/IJMS18122653

[69] Daroach M, Narang T, Saikia UN, Sachdeva N, Sendhil Kumaran M. Correlation of vitamin D and vitamin D receptor expression in patients with alopecia areata: a clinical paradigm. Int J Dermatol 2018;57:217-22. https://doi.org/10.1111/IJD.13851

[70] Georgieff MK. Iron deficiency in pregnancy. Am J Obstet Gynecol 2020;223:516-24. https://doi.org/10.1016/J.AJOG.2020.03.006

[71] Kantor J, Kessler LJ, Brooks DG, Cotsarelis G. Decreased serum ferritin is associated with alopecia in women. J Invest Dermatol 2003;121:985-8. https://doi.org/10.1046/J.1523-1747.2003.12540.X

[72] Trost LB, Bergfeld WF, Calogeras E. The diagnosis and treatment of iron deficiency and its potential relationship to hair loss. J Am Acad Dermatol 2006;54:824-44. https://doi.org/10.1016/J.JAAD.2005.11.1104

[73] Shrivastava S. Diffuse hair loss in an adult female: approach to diagnosis and management. Indian J Dermatol Venereol Leprol 2009;75:20-8. https://doi.org/10.4103/0378-6323.45215

[74] Du X, She E, Gelbart T, Truksa J, Lee P, Xia Y, Khovananth K, Mudd S, Mann N, Moresco EMY, Beutler E, Beutler B. The serine protease TMPRSS6 is required to sense iron deficiency. Science 2008;320:1088-92. https://doi.org/10.1126/SCIENCE.1157121

[75] Ju W, Ji M, Li X, Li Z, Wu G, Fu X, Yang X, Gao X. Relationship between higher serum selenium level and adverse blood lipid profile. Clin Nutr 2018;37:1512-7. https://doi.org/10.1016/J.CLNU.2017.08.025

[76] Bates JM, Spate VL, Steven Morris J, St Germain DL, Galton VA. Effects of Selenium Deficiency on Tissue Selenium Content, Deiodinase Activity, and Thyroid Hormone Economy in the Rat during Development. Endocrinology 2000;141:2490-500. https://doi.org/10.1210/ENDO.141.7.7571

[77] Sengupta A, Lichti UF, Carlson BA, Ryscavage AO, Gladyshev VN, Yuspa SH, Hatfield DL. Selenoproteins Are Essential for Proper Keratinocyte Function and Skin Development. PLoS One 2010;5:e12249. https://doi.org/10.1371/JOURNAL.PONE.0012249

[78] MacFarquhar JK, Broussard DL, Melstrom P, Hutchinson R, Wolkin A, Martin C, Burk RF, Dunn JR, Green AL, Hammond R, Schaffner W, Jones TF. Acute selenium toxicity associated with a dietary supplement. Arch Intern Med 2010;170:256. https://doi.org/10.1001/ARCHINTERNMED.2009.495

[79] Masumoto K, Nagata K, Higashi M, Nakatsuji T, Uesugi T, Takahashi Y, Nishimoto Y, Kitajima J, Hikino S, Hara T, Nakashima K, Nakashima K, Oishi R, Taguchi T. Clinical features of selenium deficiency in infants receiving long-term nutritional support. Nutrition 2007;23:782-7. https://doi.org/10.1016/J.NUT.2007.08.001

[80] Petru E, Petru C, Benedicic C, Sieja K, Talerczyk M. Re: "Selenium as an element in the treatment of ovarian cancer in women receiving chemotherapy." Gynecol Oncol 2005;96:559-60. https://doi.org/10.1016/J.YGYNO.2004.09.060

[81] Uwitonze AM, Ojeh N, Murererehe J, Atfi A, Razzaque MS. Zinc Adequacy Is Essential for the Maintenance of Optimal Oral Health. Nutrients 2020;12:949. https://doi.org/10.3390/NU12040949

[82] Ogawa Y, Kawamura T, Shimada S. Zinc and skin biology. Arch Biochem Biophys 2016;611113-9. https://doi.org/10.1016/J.ABB.2016.06.003

[83] Finner AM. Nutrition and Hair. Deficiencies and Supplements. Dermatol Clin 2013;31:167-72. https://doi.org/10.1016/j.det.2012.08.015

[84] Abdel Fattah NSA, Atef MM, Al-Qaradaghi SMQ. Evaluation of serum zinc level in patients with newly diagnosed and resistant alopecia areata. Int J Dermatol 2016;55:24-9. https://doi.org/10.1111/IJD.12769

[85] Dubey P, Thakur V, Chattopadhyay M. Role of minerals and trace elements in diabetes and insulin resistance. Nutrients 2020;12:1-17. https://doi.org/10.3390/NU12061864

[86] Robberecht H, Verlaet AAJ, Breynaert A, de Bruyne T, Hermans N. Magnesium, Iron, Zinc, Copper and Selenium Status in Attention-Deficit/Hyperactivity Disorder (ADHD). Molecules 2020;25:4440. https://doi.org/10.3390/MOLECULES25194440

[87] Dodig-Curković K, Dovhanj J, Curković M, Dodig-Radić J, Degmecić D. Uloga cinka u lijecenju hiperaktivnog poremećaja u djece. The role of zinc in the treatment of hyperactivity disorder in children. Acta Med Croatica 2009;63:307-13.

[88] Kil MS, Kim CW, Kim SS. Analysis of serum zinc and copper concentrations in hair loss. Ann Dermatol 2013;25:405-9. https://doi.org/10.5021/AD.2013.25.4.405

[89] Davis C, Bryan J, Hodgson J, Murphy K. Definition of the Mediterranean Diet; A Literature Review. Nutrients 2015:7:9139-53. https://doi.org/10.3390/nu7115459

[90] Ventriglio A, Sancassiani F, Contu MP, Latorre M, Di Salvatore M, Fornaro M, Bhugra D. Mediterranean Diet and its Benefits on Health and Mental Health: A Literature Review. Clin Pract Epidemiol Ment Health 2020;16:156-64. https://doi.org/10.2174/1745017902016010156

[91] Cannataro R, Fazio A, La Torre C, Cristina Caroleo M, Cione E, Albertini C. Antioxidants Polyphenols in the Mediterranean Diet: From Dietary Sources to microRNA Modulation. Antioxidants 2021;10:328. https://doi.org/10.3390/antiox10020328

[92] Amleh A, Sharif J, Jeronimo C, López-Camarillo C, Á C-R, Carlos-Reyes Á, Sullivan López-González J, Meneses-Flores M, Gallardo-Rincón D, Ruíz-García E, Marchat LA, Astudillo-de la Vega H, Hernández de la Cruz ON. Dietary Compounds as Epigenetic Modulating Agents in Cancer. Front Genet 2019;10:79. https://doi.org/10.3389/fgene.2019.00079

[93] Ganesan K, Xu B. A critical review on polyphenols and health benefits of black soybeans. Nutrients 2017;9:455. https://doi.org/10.3390/nu9050455

[94] Bonaccio M, Pounis G, Cerletti C, Donati MB, Iacoviello L, de Gaetano G. Mediterranean diet, dietary polyphenols and low grade inflammation: results from the MOLI-SANI study. Br J Clin Pharmacol 2017;83:107-13. https://doi.org/10.1111/BCP.12924

[95] Bassino E, Gasparri F, Munaron L. Protective role of nutritional plants containing flavonoids in hair follicle disruption: a review. Int J Mol Sci 2020;21:523. https://doi.org/10.3390/IJMS21020523

[96] Bassino E, Gasparri F, Munaron L. Protective role of nutritional plants containing flavonoids in hair follicle disruption: a review. Int J Mol Sci 2020;21:523. https://doi.org/10.3390/IJMS21020523

[97] Fortes C, Mastroeni S, Mannooranparampil T, Abeni D, Panebianco A. Mediterranean diet: fresh herbs and fresh vegetables decrease the risk of Androgenetic Alopecia in males. Arch Dermatol Res 2018;31071-6. https://doi.org/10.1007/s00403-017-1799-z

[98] Murata K, Noguchi K, Kondo M, Onishi M, Watanabe N, Okamura K, Matsuda H. Promotion of hair growth by Rosmarinus officinalis leaf extract. Phytother Res 2013;27:212-7. https://doi.org/10.1002/PTR.4712

[99] Malvezzi De Macedo L, Mendes Dos Santos É, Militão L, Tundisi LL, Ataide JA, Barbosa Souto E, Gava Mazzola P. Plants Rosemary (Rosmarinus officinalis L., syn Salvia rosmarinus Spenn.) and Its Topical Applications: A Review. Plants (Basel) 2020;9:651. https://doi.org/10.3390/plants9050651

[100] Wallace TC. Health effects of coconut oil-a narrative review of current evidence. J Am Coll Nutr 2019;38:97-107. https://doi.org/10.1080/07315724.2018.1497562

[101] Hamel AF, Menard MT, Novak MA. Fatty Acid supplements improve hair coat condition in rhesus macaques. J Med Primatol 2017;46:248-51. https://doi.org/10.1111/jmp.12271

[102] Kong J, Qiang W, Jiang J, Hu X, Chen Y, Guo YX, Liu H, Sun S, Gao H, Zhang Y, Gao Y, Liu X, Liu X, Li H. Safflower oil body nanoparticles deliver hFGF10 to hair follicles and reduce microinflammation to accelerate hair regeneration in androgenetic alopecia. Int J Pharm 2022;616121537. https://doi.org/10.1016/J.IJPHARM.2022.121537

[103] Lourith N, Kanlayavattanakul M, Chaikul P. Para rubber seed oil: The safe and efficient bio-material for hair loss treatment. J Cosmet Dermatol 2021;20:2160-7. https://doi.org/10.1111/JOCD.13843

[104] Tenore GC, Caruso D, Buonomo G, D'Avino M, Santamaria R, Irace C, Piccolo M, Maisto M, Novellino E. Annurca apple nutraceutical formulation enhances keratin expression in a human model of skin and promotes hair growth and tropism in a randomized clinical trial. J Med Food 2018;21:90-103. https://doi.org/10.1089/jmf.2017.0016

[105] Zhang Y, Ni C, Huang Y, Tang Y, Yang K, Shi X, Zhang Y, Li Z, Wang J', Zhu Y, Li H, Ma Y, Lin J, Wang J, Liu Q, Wu W. Hair growth-promoting effect of resveratrol in mice, human hair follicles and dermal papilla cells. Clin Cosmet Investig Dermatol 2021;14:1805-14. https://doi.org/ 10.2147/CCID.S335963

[106] Katoulis AC, Liakou AI, Koumaki D, Vakirlis E, Tsantes AG, Mortaki D, Bozi E, Ioannides D. A randomized, single-blinded, vehicle-controlled study of a topical active blend in the treatment of androgenetic alopecia. Dermatol Ther 2020;33:E13734. https://doi.org/10.1111/DTH.13734

[107] Iltaf J, Noreen S, Rehman MFU, Ghumman SA, Batool F, Mehdi M, Hasan S, Ijaz B, Akram MS, Butt H. Ficus benghalensis as Potential Inhibitor of 5α-Reductase for Hair Growth Promotion: In Vitro, In Silico, and In Vivo Evaluation. Front Pharmacol 2021;12:774583. https://doi.org10.3389/fphar.2021.774583

Correspondence: Munis Dundar, Department of Medical Genetics, Faculty of Medicine, Erciyes University, 38039, Kayseri, Turkey. E-mail: dundar@erciyes.edu.tr

How to cite this article: Gokce N, Basgoz N, Kenanoglu S, Akalin H, Ozkul Y, Ergoren MC, Beccari T, Bertelli M, Dundar M. An overview of the genetic aspects of hair loss and its connection with nutrition. J Prev Med Hyg 2022;63(suppl.3):E228-E238. https://doi.org/10.15167/2421-4248/jpmh2022.63.2S3.2765

© Copyright by Pacini Editore Srl, Pisa, Italy

This is an open access article distributed in accordance with the CC-BY-NC-ND (Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International) license. The article can be used by giving appropriate credit and mentioning the license, but only for non-commercial purposes and only in the original version. For further information: https://creativecommons.org/licenses/by-nc-nd/4.0/deed.en

PRODUCT REVIEW

Human Vaccines & Immunotherapeutics 11:4, 908–914; April 2015; © 2015 Taylor & Francis Group, LLC

The development of a recombinant hepatitis E vaccine HEV 239

Shao-Wei Li[1,2,3], Qinjian Zhao[1,2], Ting Wu[1,*], Shu Chen[1,2], Jun Zhang[1,2], and Ning-Shao Xia[1,2]

[1]National Institute of Diagnostics and Vaccine Development in Infectious Diseases; Xiamen University; Xiamen, PR China; [2]State Key Laboratory of Molecular Vaccinology and Molecular Diagnostics; School of Public Health; Xiamen University; Xiamen, PR China; [3]School of Life Science; Xiamen University; Xiamen, PR China

Keywords: efficacy, Hecolin, Hepatitis E virus, HEV 239, vaccine

Hepatitis E virus (HEV) infection is one of the main causes of acute hepatitis worldwide. A recombinant hepatitis E vaccine, HEV 239, has been licensed in China for immunizing adults of 16 y old and above. The vaccine antigen contains pORF2 aa 368 - 606 of the HEV genotype 1 expressed in *E. coli*. The quality of the vaccine is controlled through a combination of biophysical, biochemical and immunochemical methods. The vaccine is well tolerated in adults. The efficacy of the HEV 239 vaccine against symptomatic and asymptomatic infection had been proven to be high during a Phase III clinical trial and long-term follow up. The safety and efficacy of HEV 239 vaccine in certain high-risk populations remains to be further investigated.

Introduction

Hepatitis E (HE) is caused by infection with the hepatitis E virus (HEV). The awareness of this disease has increased a lot in recent years.[1] HEV infection remains a serious threat to life and productivity in the developing world. In the developed countries, an increasing number of autochthonous HE cases have been reported, suggesting that this threat is not limited to developing regions. Patients with underlying liver disease and patients who are immunosuppressed or immunocompromised are at particular risk.[2]

HEV is a non-enveloped, single-stranded, positive-sense RNA virus. It is the only member of the genus *Hepevirus* in the family *Hepeviridae*.[3] Four genotypes of HEV are associated with human diseases: genotypes 1 and 2 infect only humans, while genotypes 3 and 4 infect many animal species in addition to humans.[3,4] All 4 genotypes are characterized as one serotype, making the development of a univalent hepatitis E vaccine reasonable. Two candidate hepatitis E vaccines have been evaluated in clinical trials. The first of these, rHEV,[5] was developed by GlaxoSmithKine (GSK, Rixensart, Belgium). The other, HEV 239 (trade name Hecolin®),[6-8] was developed by Innovax (Xiamen, China) (**Table 1**). HEV 239 was launched in China in 2012. In this review, we focus on the development of the HEV 239 vaccine.

Structural basis for the HEV 239 vaccine

The HEV genome consists of 3 open reading frames (ORFs). ORF2 encodes a single structural protein (pORF2) for the HEV capsid. Structurally, the HEV capsid is arranged as a $T = 3$ icosahedral lattice, which assembles with the coordination of the HEV RNA genome.[9] The approximately 110 amino acid residues (aa) of the N-terminal region are responsible for RNA binding, given the positively charged arginine-clustered region here. With truncations in the 111 N-terminal aa and the 53 C-terminal aa, the protein can self-assemble into a $T = 1$ icosahedral VLP via a baculovirus / insect cell expression system.[10] The VLP form of genotypes 3 and 4 can be crystallized to allow the structural determination at a resolution of 3.5 Å.[11,12] The analysis of these structures has revealed 3 functional domains of pORF2, including the S domain (aa 118–313), which contributes to the 5-fold interaction and the formation of the basal shell; the P1 domain (aa 314–453), which is important for the 3-fold interaction; and the P2 domain (aa 454–606), which is involved in the 2-fold interaction and is responsible for host interactions.[11,12] The P2 domain, also called the E2s domain, was further analyzed, and its high-resolution crystal structure was determined. The crystal structure showed that two E2s form a tight homodimer through the strong hydrophobic interaction maintained by side chains of residues Val503, Trp548, Thr552, Ala555, Tyr557, Tyr561, Val598 and Val600 of both subunits of the dimer.[13] There is a unique groove region located on the most protrusive moiety along the 2-fold axis. This groove region was verified to be involved in virus-host interactions using site-directed mutagenesis and the binding of neutralizing mAbs.[13] The structure of the E2s domain in a complex with neutralizing mAb 8C11 further pinpointed the neutralization sites at Asp496-Thr499, Val510-Leu514, and Asn573-Arg578. Mutation and cell-model assays identified Arg512 as the most crucial residue for 8C11 interaction with and neutralization of HEV.[14] It should be noted that the epitope recognized by 8C11 represents a major immune dominant epitope. As such, it is capable of blocking naturally acquired anti-HEV from binding with the E2 antigen.[15]

HEV 239 was designed as ORF2 aa368–606 fragment from a further N-terminal truncation of $T = 1$ VLP and was adapted to be expressed with a high efficiency in *E.coli*.[16] During the development of the HEV 239 vaccine, some historic constructs based on antigenicity and immunogenicity were produced, including E2s (aa459–606), E2(aa394-606), HEV239(aa368-606) and T=1 VLP (112-607). E2s forms a dimer with 8C11 binding activity (**Fig. 1**).

*Correspondence to: Ting Wu; Email: wuting@xmu.edu.cn

Submitted: 11/09/2014; Revised: 11/30/2014; Accepted: 12/09/2014
http://dx.doi.org/10.1080/21645515.2015.1008870

Table 1. Characteristics of the 2 existing hepatitis E vaccines that have been evaluated in clinical trials

	rHEV	HEV 239 (Hecolin®)
Manufacturer	GlaxoSmithKine, Belgium	Innovax, China
Origin	HEV genotype 1	HEV genotype 1
Expression system	Baculavirus	E. coli
Gene location	ORF2 aa 112–607	ORF2 aa 368–606
Antigen content	20 μg	30 μg
Adjuvant	Aluminum hydroxide	Aluminum hydroxide
Immunization schedule	0, 1, and 6 month	0, 1, and 6 month
Sample size of the efficacy trial	2 000	112 604
Participants enrolled	Mostly male (99.6%), aged from 18–62 (mean age 25.2)	Female and male, aged from 16–65 (mean age 44.1)
Efficacy against hepatitis E in the PPS cohort	96% (95% CI, 86–99)	100% (95% CI, 72–100)
Efficacy against hepatitis E in the ITT cohort	89% (95% CI, 77–94)	96% (95% CI, 66–99)
Development stage	Phase II clinical trial finished	Licensed in China
Advantages and disadvantages	Efficacy in women and in some special population remains to be determined	Efficacy in some special population remains to be determined

PPS: per-protocol set; ITT: intention to treat.

E2 (aa394–606) forms a hexamer with considerable antigenicity and marked immunogenicity in rhesus monkeys. The N-terminal 26 aa extension of E2, called HEV 239, self-assembles into shrunken version of the virus-like particles of the ~20 nm diameter. These particulate particles verified with plausible antigenicity and immunogenicity.[16, 17] Although HEV 239 (aa 368–606) protein is N-terminally shorter than T = 1 VLP (aa 112–606) in aa sequence, but they share the structural feature of outmost moiety of viral capsid (**Fig. 1**), which, E2s, or called P2 domain, is responsible for virus-host interaction and subsequently accounts for antigenicity and immunogenicity in vaccine design. Compared with the complex expression in insect cell, production in E.coli is easy to handle and scale-up. HEV 239 can self-assemble into particles in vitro. However, T = 1 VLP was only found forming particle in insect cell (rHEV vaccine,[5] GSK). In clinical trials, HEV 239 and T = 1 VLP conferred comparable hepatitis E protection in vaccine recipients.[5,7] In terms of nanoscale molecules used for antigen presenting cell processes, VLPs were verified to be the key characteristics for immune recognition.[17,18] Based on this, successful recombinant vaccines for hepatitis B and human papillomavirus were brought to market.[17,18] Thus, HEV 239 was considered for further preclinical study as a HEV vaccine.

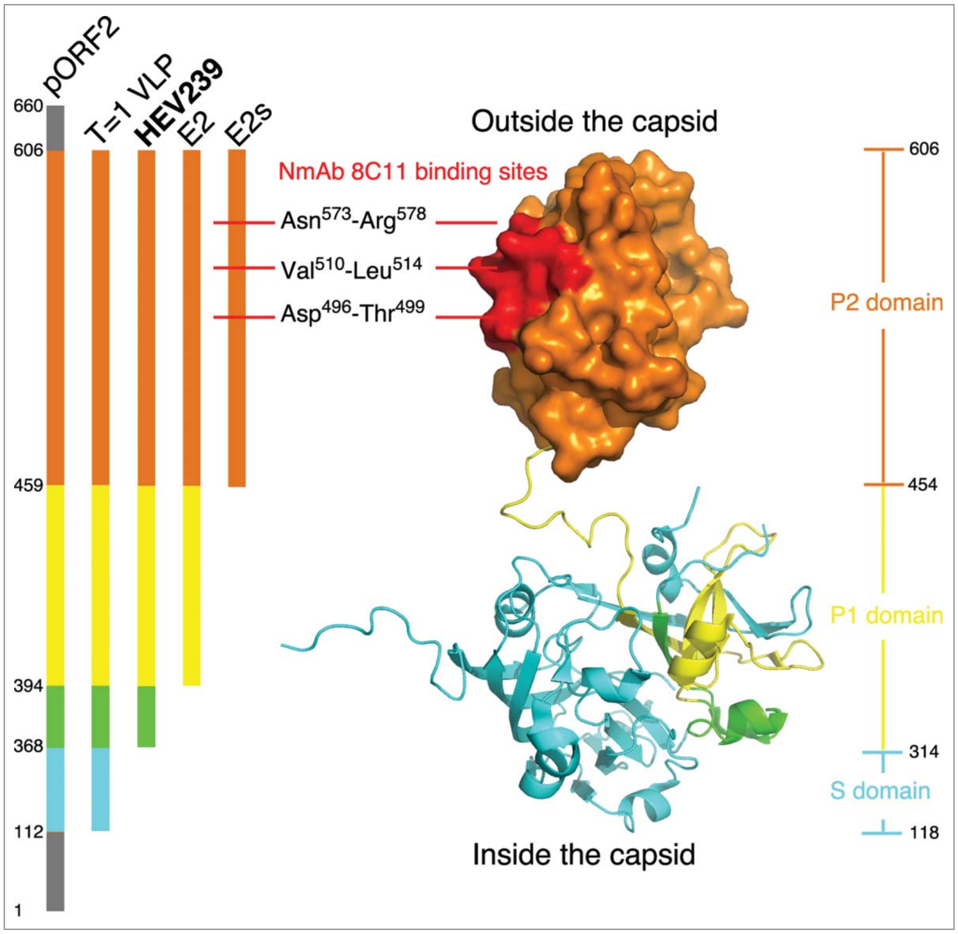


Figure 1. Structural interpretation for HEV 239 vaccine. The structure of HEV ORF2 aa112–606 is rendering in surface mode (P2 domain) and cartoon mode (P1 and S domain), colored by key regions for several HEV historic constructs (left panel), i.e., aa112–367 in cyan, aa368–393 in green, aa394–458 in yellow and aa459–606 in orange. To elucidate the spatial configuration, 3 crystal structures were in superimposition to demonstrate major neutralization sites (Neutralizing mAb 8C11 binding sites colored in red, by PDB (protein database) no. 3RKD), outmost protrusion domain (E2s domain in orange surface, by PDB no. 3GGQ) and P1&S domains (contributing for virus capsid assembly, by PDB no. 3HAG). The figure was prepared by the program PyMol.

The preclinical development of HEV 239 vaccine

HEV 239 was over expressed under the regulatory of T7 promoter in *E. coli*, the gene was cloned into a non-fusion plasmid vector pTO-T7 and subsequent expression was induced by IPTG. HEV 239 forms inclusion bodies in bacteria. These inclusion bodies were harvested by liquid-solid separation and panwashing with Triton-X100 before chaotropic agent denaturation. After harvesting, the resulting ~70% pure HEV 239 inclusion bodies were dissolved in 4 M urea. The HEV 239 protein was further purified to a final purity of 95–98% by 2 polishing chromatography steps and then subjected to particle reassembly with a scalable tangential flow technique, using a membrane cassette with a cutoff size of 30kD.[16] The reassembly condition was optimized for a robust range of pH, ion-strength and temperature. Essentially, HEV 239 can self-assemble in 4 M urea in the presence of salt, which facilitates the assembly control during this process.[19] Self-assembled HEV 239 was further fractionalized using capture chromatography, and the component with an average radius of ~10 nm was collected and absorbed onto an aluminum adjuvant.[16]

The pilot process scale was developed in 10 L fermentation batches, which generated several thousands of vials of vaccine for investigational new drug (IND) purposes. Briefly, the conditions for the serial Essentially, HEV 239 processing were gauged using a variety of biochemical and biophysical approaches, including SDS-PAGE, Western blotting, DLS (Dynamic Light Scattering), SEC HPLC (Size-Exclusion Chromatography), AFM (Atom Force Microscopy), AUC (Analytical Ultracentrifugation), DSC (Differential Scanning Calorimetry) and TEM (Transmission Electronic Microscopy).[20] Antigenicity was monitored using a sandwich ELISA with mAbs 8C11 and 8H3. This mAb couple identified the E2s moiety of HEV 239 on the groove region of the E2s crystal structure.[13] As a rare event in virology, the binding of 8C11 with an antigen will significantly enhance the binding affinity of another mAb 8H3 by 2 log in SPR (Surface Plasmon Resonance) assay. The synergistic action of this mAb couple also was discernible in authentic HEV.[15] All residual contaminants were evaluated and quantified under the regulatory limits issued by the China Food and Drug Administration (CFDA). The immunogenicity of the formulated HEV 239 was measured *in vivo* with the ED_{50} (50% effective dose) test in mice and with the anti-HEV levels in rhesus monkeys. The ED_{50} value for HEV 239 in mice was approximately 0.05 μg, demonstrating that HEV 239 has a high level of immunogenicity. Indeed, the immunogenicity of HEV 239 was 200 times greater than that of the E2 protein.[16,21] The protection efficacy of HEV 239 was primarily evaluated in the rhesus monkey model. Two doses of 5 μg, 10 μg or 20 μg HEV 239 were administered 4 weeks apart. The induced antibody response peaked at week 7 in all 3 dosage groups with comparable antibody titers of around 1384 IU. The anti-HEV levels were measured by ELISA assay and calibrated against a World Health Organization (WHO) reference standard for anti-HEV (National Institute for Biological Standards and Control, NIBSC catalog no. 95/584), which was obtained from an recovered hepatitis E patients with an arbitrary assigned value of 100 IU.[16] The monkeys were then challenged with 10^4 or 10^7 genome equivalences of homologous HEV (genotype 1) or heterologous HEV (genotype 4). All of the HEV 239 vaccinated monkeys ($n = 24$) were completely protected from symptomatic hepatitis E and were also completely protected from subclinical infection by 10^4 HEV ($n = 12$). When challenged with 10^7 HEV, 9 of the 12 monkeys (75%) were protected from subclinical infection. Among the 3 monkeys with breakthrough HEV infection, 2 monkeys (with anti-HEV level of 368 IU and 196 IU) shed virus during the first week post infection and the third monkey (anti-HEV level 961 IU) had detectable HEV RNA in feces during week 3. The vaccine protection efficacy against homologous HEV (genotype 1) or heterologous HEV (genotype 4) infection was similar.[16] After the further safety evaluation of HEV 239 in animals, which including intramuscular irritation in rabbits, acute toxicity testing in KM (Kunming) mice, deformity testing in gravid white KM mice and repetitive immune toxicity testing in SD (Sprague dawley) rats, the HEV 239 vaccine was approved by the CFDA for further clinical trials.[8,22]

Assessment the HEV 239 vaccine critical quality attributes (CQAs)

The product consistency of the HEV 239[8,20,23] vaccine was demonstrated using a combination of biophysical, biochemical and immunochemical methods. Results from these methods demonstrated that different vaccine batches have comparable antigen characteristics or CQAs.

The presence of clinically relevant epitopes can be used as a surrogate marker for vaccine potentcy.[17,24,25] A large panel of monoclonal antibodies (mAbs) against the HEV capsid protein were developed to improve the understanding of HEV virology and to analyze the CQAs of the vaccine during bioprocessing and in the lot release and stability testing.[14,15,20,26,27] Key epitopes of the particulate antigen were analyzed using a panel of murine anti-HEV mAbs against different epitopes, analogous to the approach used for other vaccines.[24,25,28-31] Epitope mapping using cross-inhibitory ELISA, surface plasmon resonance biosensor chip based binding assay, and sandwich ELISA using HEV 239 specific mAbs were used to assure vaccine potency.[20,26,27] Overall, our findings showed that the antigen production process is robust and scalable in Hecolin® manufacturing.

A successful and robust scale-up (i.e., 50-fold) in the production process (1.0 L, 10 L and 50 L) was demonstrated.[20] Multiple HEV 239 batches were produced at a commercial scale (50 L) and were characterized using several different methods (**Table 2**).

Biochemical methods

To verify the consistency of the primary amino acid sequence, analyses were carried out to determine the antigen purity, antigen integrity, molecular weight, pI and protein sequence.[20,27] The most prominent band on the SDS-PAGE gel showed an apparent molecular weight of 29.4 kDa. All three lots exhibited similar MWs by MALDI-TOF-MS, with a mean MW of 25,561.5 ± 1.5 Da (**Table 1**), which is consistent with the theoretical MW of HEV 239. In addition, LC-MS-based protein mapping with tryptic digestion was used to determine the peptide sequence.

Table 2. Analytical toolbox for the comprehensive characterization of a recombinant HEV 239 antigen to demonstrate the process reproducibility and product consistency during Hecolin® production.[20,26,27]

Method		
Biochemical methods	Characteristics	Parameters of HEV 239 or Notes (from multiple lots, n = 6)
SDS-PAGE	Purity, integrity, molecular weight (kDa)	29.32 ± 0 .08
MALDI-TOF MS	Molecular weight (Da)	25,561.64 ± 1 .37
LC-MS	Peptide mapping (coverage, %)	100 (3 lots)
icIEF	pI, whole column detection capillary isoelectric focusing	6.55 ± 0 .03
Biophysical methods		
HPSEC	Size (as reflected by retention time, min)[b]	14.02 ± 0 .08
TEM	Size, morphology	20–30 nm particles with certain degree of irregularity
DLS	Hydrodynamic diameter (nm)	28.92 ± 0 .20
AUC	Sedimentation coefficient (S)	21.42 ± 0 .13
DSC	Thermal stability (°C)	75.58 ± 0 .10
Cloud point	Aggregation propensity (°C)	71.83 ± 0 .16
CD	Secondary structure	The secondary structure, including α-helices and β-sheets, is
UV	Trp/Tyr absorption	consistent with the UV spectrum of a globular protein with a maximum absorption wavelength at 280 nm
Immunochemical methods		
SPR With mAbs	affinity to mAbs	The binding activity to 5 mAbs for different batches of aqueous products with the RU(Ag)/ RU(Ab)
Sandwich ELISA[d]	Antigenicity	8C11 (capture) and 8H3 (detection) to analyze the consistency of the antigenicity among different batches of HEV 239 aqueous products
Immunological assessment		
Mouse potency	Immunogenicity ED_{50} (μg)	0.025–0.060

Note: 1. For product lot release of prefilled syringe, mouse potency assay is used for lot release (spec < 1.0 μg). Lot release of HEV 239 purity (spec > 95.0%) and molecular weight (spec 27.5 ± 2.6 kDa) by SDS-PAGE was performed on aqueous product batches.

Two. The reported values were calculated from the results on 6 manufacturing lots. These values are representative on process performance since similar values were reported for sizing using DLS and antigenicity test using the results from over 20 lots.[20]

The overall amino acid sequence was confirmed with 100% coverage for all 3 batches.[20]

Biophysical methods

Four different sizing methods (TEM, HPSEC, AUC and DLS) were used to monitor the particle size consistency (**Table 1**). The HEV 239-based VLPs were found to be 20–30 nm in diameter by TEM, with a certain degree of irregularity and heterogeneity.[20] HPSEC analysis resulted in nearly identical retention time for all 6 lots of HEV 239 aqueous products.[20] The AUC profiles demonstrated highly similar sedimentation properties for the 6 different lots, all of which peaked at approximately 21.5 s.[20]

Both circular dichroism and UV spectroscopy methods verified the consistency in the secondary and tertiary protein structure among the different HEV 239 lots.[20]

The thermal unfolding of the VLP antigen was analyzed using DSC by monitoring the T_m values of all HEV 239 lots, all of which ranged from ~75–76°C. These results showed good cross lot product consistency as reflected by the thermal stability of the recombinant protein antigen during heat-induced unfolding.[20,27] In addition, the propensity for the protein to aggregate under thermal stress was assessed using cloud point analysis. The heat-induced unfolding or aggregation profiles of the VLPs were highly comparable among all 6 lots, as determined by both DSC and cloud point (**Table 2**).

Immunochemical methods

The *in vitro* antigenicity of an antigen, as probed with different mAbs, correlates in general to the *in vivo* immunogenicity of a vaccine.[30-33] Therefore, a one-site binding and label-free assay (i.e., surface plasmon resonance or SPR)[34-36] and a 2-site binding assay (i.e., a sandwich ELISA)[30,31,33] were carried out to evaluate the antibody binding activity of the HEV 239 antigen.

The label-free and real time sensor chip-based SPR method was used with 5 mAbs recognizing different epitopes (mouse IgGs-8C11, 8H3, 13D8, 12A10 and 16D7) to assess the immunoreactivity of the different batches.[20] A sandwich ELISA was implemented in a manufacturing setting with the desired robustness and reproducibility to show product consistency.[20] A well-characterized neutralizing and protective mAb, 8C11, was selected as the capture antibody,[26] and 8H3 labeled with HRP was used as the detection antibody.

Immunological assessment

To measure the *in vivo* potency of the vaccine immunogenicity, 6 batches of HEV 239 were evaluated by measuring the ED_{50} in a mouse model after absorption onto an amorphous aluminum-based adjuvant. All six batches had comparable ED_{50} values in the range of 0.025–0.060 μg (**Table 2**). Two of the assays used here, the SDS-PAGE and the mouse potency assay (ED_{50} determination), are used for routine lot release testing on the final vaccine formulation and filled product. Based on assay results on recovered antigen, full antigen recovery was observed post adjuvant dissolution with no alterations on antigen characteristics.[27]

Clinical development of HEV 239

All pre-approval clinical trials were carried out in China in accordance with guidelines and regulations issued by the CFDA and the provisions of the Declaration of Helsinki.

Efficacy of HEV 239

A randomized, double-blind, placebo-controlled, single-center Phase III clinical trial was conducted to assess the efficacy of the HEV 239 vaccine against symptomatic hepatitis E. Participants aged 16–65 ($n = 112\ 604$) were randomly assigned to receive 30 μg HEV 239 (absorbed to 0.8 mg of aluminum hydroxide suspended in 0.5 mL of buffered saline) or 5 μg of the hepatitis B vaccine control intramuscularly at 0, 1, and 6 months. The patients were followed for 19 months,[7] and symptomatic hepatitis E cases were identified though a sensitive surveillance system containing 205 local sentinels at the site. The efficacy was demonstrated to be 100% (95% CI, 72–100) in participants receiving all 3 doses of the HEV 239 vaccine or the control vaccine (PPS, per-protocol set) during the following 12 months, starting at 30 d after the receipt of the final dose (7 m–18 m). In the PPS cohorts, no HEV 239 vaccine recipients contracted hepatitis E, whereas 15 hepatitis E cases developed in the control group during the same year. For those who received at least one dose (ITT, intention to treat cohort), the efficacy was 96% (95% CI, 66–99) during the first 18 months.

In an extended efficacy study,[37] all participants in the Phase III trial were followed for another 3 y using the same local hepatitis surveillance system used during the Phase III trial. The long-term efficacy of HEV 239 for up to 4.5 y was demonstrated to be 93% (95% CI, 79 - 98) in the PPS cohort and 87% (95% CI, 71 - 94) in the ITT cohort. Due to the large cohorts of the Phase III trial, only participants in the immunogenicity subset were tested for anti-HEV IgG level at month 7, and as there had been no breakthrough hepatitis E cases in this immunogenicity subset with known anti-HEV level at month 7 until now, the protective level of anti-HEV can not be calculated.

As expected, 90% of the 29 HEV isolates from the identified hepatitis E patients were of genotype 4, which is the predominant HEV genotype circulating in China. The high efficacy of HEV 239, which is derived from HEV genotype 1, against HEV genotype 4 infection supported the concept that a hepatitis E vaccine containing one genotype might protect against infection from different HEV genotypes, which are capable of infecting mammals and share the same serotype.

The HEV 239 vaccine has also been shown to be effective in the prevention of subclinical HEV infection.[6,38] Asymptomatic HEV infection was identified by comparing paired serum samples from the same participant collected at 12 month intervals, starting 30 d after administration of the final dose when the antibody levels were supposed to have peaked. If the anti-HEV IgG level rose from negative to 0.154 Wu/mL (twice of the cut-off level) or increased by 4-fold or greater, patients were considered to have asymptomatic infection. The relative risk of HEV infection in the vaccine recipients decreased with increases in the anti-HEV IgG titer at month 7. Even marginal concentration of antibodies significantly lowered the infection risk. This might explain why the efficacy against subclinical HEV infection is similar in the PPS (79%) and ITT cohorts (77%).[38] Similar efficacy rates had been demonstrated in the Phase II trial, regardless of whether 3 doses (given at 0, 1 and 6 m) or 2 doses (given at 0, 1 and 6 m) of vaccine were given. In both groups, the tested vaccine contains only 20 μg of the HEV 239 antigen.

The administration of 2 doses of HEV 239 given at months 0 and 1 yielded efficacy rates of 100% (95% CI, 9–100) against hepatitis E during months 2 to 6, before the third shot.[7] These data justify the possibility of emergency vaccination during epidemics. Similarly, 2 doses of HEV 239 vaccine given at one month intervals provided satisfying protection against subclinical HEV infection for at least the following 5 months.[6]

Immunogenicity of HEV 239

HEV 239 administration induced a vigorous anti-HEV IgG response in vaccine recipients, regardless of their initial antibody status. Of the baseline seronegative participants, 99.9% had seroconverted one month after the full vaccination course (7 m), and 87% remained seropositive at month 55. The geometric mean concentration (GMC) of antibody dropped quickly from a peak of 14.96 Wu/ml at month 7, which is higher than the level induced by natural HEV infection (0.6 Wu/ml) but lower than that induced after symptomatic hepatitis E disease (80.9 Wu/ml), to 1.47 Wu/ml for the first year.[37,38] The concentration then slowly decreased to 0.27 Wu/ml by month 55. Although the antibody decayed to a relative low level at this time point, the long-term efficacy until month 54 is still as high as 93% (95% CI, 79–98) in the PPS cohort, the need and timing for a booster dose of HEV 239 remains to be further investigated. For participants with pre-existing antibodies, the antibody response is more robust and sustained, with little difference between those who received 1, 2 or 3 doses of HEV 239 vaccine. This finding suggests that people with pre-existing HEV antibodies would benefit from even a single dose of the HEV 239 vaccine.[7,37] In a HBsAg positive cohort, the immunogenicity of HEV 239 was found to be similar to that of the general population.[39]

Anti-HEV IgG levels induced by different dosage and immunization schedules of the HEV 239 vaccine were also evaluated in the Phase II trial, in which only subjects who were anti-HEV seronegative at baseline were enrolled. All participants who received 3 doses of the HEV 239 vaccine seroconverted by month 7, and the level of immune response increased as the dosage increase from 10 μg to 40 μg. The difference in the geometric mean antibody concentration (GMC) was significant between the 30 μg and the other higher doses ($p < 0.05$). However, the difference in the GMC between 20, 30 and 40 μg was not statistically significant. The study found that the 3 doses course was more potent than the 2 doses course.[6]

Safety of HEV 239

The data from all clinical trials suggest that the HEV 239 vaccine is well tolerated.[40] In the Phase I trial, 44 healthy Chinese adults were enrolled, and all were anti-HEV IgG seronegative with normal serum biochemistry values. Then, 20 μg of the HEV 239 vaccine was administered at month 0 and 1. All participants were followed for 60 d after the first dose, and their serum biochemistry values were analyzed at day 0 and 60. No serious adverse events were reported. Furthermore, the serum biochemistry values of the participants, including fibrinogen (FBG), urea nitrogen (UREA), glucose (GLU), alanine aminotransferase (ALT), showed no abnormal changes from their baseline values by day 60.[8]

The Phase III clinical trial had 112,604 participants. In this trial, data obtained from the "reactogenicity subset" (n = 2645) within the first 72 hours following each dose showed the occurrence of more frequent local adverse events in the HEV 239 group (13.5%) compared to the control hepatitis B vaccine group (7.1%). This might be partially due to the higher antigen dose in the HEV 239 vaccine (30 μg) than the hepatitis B vaccine (5 μg). Local adverse events of level 3 or higher, solicited systemic adverse events and unsolicited adverse events occurred at similar rates between the study groups. Serious adverse event (SAE) data were collected for all 112,604 participants and categorized according to the Medical Dictionary for Regulatory Activities (MedDRA). During the Phase III trial and the extended 4.5-year follow-up period, a comparable number of participants in both the vaccine and placebo groups experienced SAEs. None of the SAEs were believed to be related to the vaccination.[7,37]

There were 37 women in the HEV 239 vaccine group and 31 women in the control group who were inadvertently administered the vaccine during pregnancy.[41] The rates of adverse events observed in the pregnant women were similar to those observed in matched non-pregnant women. The weights, body lengths and gestational ages of the newborns were comparable in both groups. Although this sample size is small and further study is needed, these data preliminarily suggest that the HEV 239 vaccine is well-tolerated in pregnant women. The safety of HEV 239 was also evaluated in HBsAg positive persons, and the resulting data are reassuring.[39]

Future Challenges

To date, Hecolin® is only available in the Chinese market and been approved for use in people older than 16 years, it is indicated for vaccinating individuals at high risk of HEV infection, such as those involved in animal husbandry, students, members of the armed forces, food-handlers, women of childbearing age, and travelers to endemic regions. The Chinese CDC has an online reporting system to collect post-marketing adverse reactions/events, and this body has not identified any safety concerns with the vaccine. Although the existing 4.5 y of safety and efficacy data for Hecolin® are reassuring, data has not yet been collected in some special populations, including as pediatric subjects (< 16 y of age), the elderly (> 65 y of age), pregnant women, persons with underlying liver diseases or those who are immunosuppressed. These groups face a greater threat of HEV infection and are therefore in more urgent need of this vaccine.

On the other hand, although epidemics of hepatitis E with thousands of cases and hundreds or tens of deaths occurred now and then in low-resource countries, HEV 239 vaccine are not available to these people until it be prequalified by WHO. The deep gap between the vaccine and those who are in urgent need of this vaccine still waits to be filled.[42]

In conclusion, the availability of Hecolin® represents an important milestone in the control and prevention of HEV. However, there is still long way to go.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Funding

The work was supported by National Major Scientific and Technological Special Project for "Significant New Drugs Development" (2013ZX09101017), National Natural Science Foundation of China (81373061), National High-tech Research & Development Program (863 Program) (2012AA02A408).

References

1. Kamar N, Bendall R, Legrand-Abravanel F, Xia NS, Ijaz S, Izopet J, Dalton HR. Hepatitis E. Lancet 2012; 379(9835):2477-88; PMID:22549046; http://dx.doi.org/10.1016/S0140-6736(11)61849-7
2. Krain LJ, Nelson KE, Labrique AB. Host immune status and response to hepatitis E virus infection. Clin Microbiol Rev 2014; 27(1):139-65; PMID:24396140; http://dx.doi.org/10.1128/CMR.00062-13
3. Purcell RH, Emerson SU. Hepatitis E: an emerging awareness of an old disease. J Hepatol 2008; 48 (3):494-503; PMID:18192058; http://dx.doi.org/10.1016/j.jhep.2007.12.008
4. Huang FF, Sun ZF, Emerson SU, Purcell RH, Shivaprasad HL, Pierson FW, Toth TE, Meng XJ. Determination and analysis of the complete genomic sequence of avian hepatitis E virus (avian HEV) and attempts to infect rhesus monkeys with avian HEV. J Gen Virol 2004; 85(Pt 6):1609-18; PMID:15166445; http://dx.doi.org/10.1099/vir.0.79841-0
5. Shrestha MP, Scott RM, Joshi DM, Mammen MP, Jr., Thapa GB, Thapa N, Myint KS, Fourneau M, Kuschner RA, Shrestha SK, et al. Safety and efficacy of a recombinant hepatitis E vaccine. N Engl J Med 2007; 356(9):895-903; PMID:17329696; http://dx.doi.org/10.1056/NEJMoa061847
6. Zhang J, Liu CB, Li RC, Li YM, Zheng YJ, Li YP, Luo D, Pan BB, Nong Y, Ge SX, et al. Randomized-controlled phase II clinical trial of a bacterially expressed recombinant hepatitis E vaccine. Vaccine 2009; 27 (12):1869-74; PMID:19168109; http://dx.doi.org/10.1016/j.vaccine.2008.12.061
7. Zhu FC, Zhang J, Zhang XF, Zhou C, Wang ZZ, Huang SJ, Wang H, Yang CL, Jiang HM, Cai JP, et al. Efficacy and safety of a recombinant hepatitis E vaccine in healthy adults: a large-scale, randomised, double-blind placebo-controlled, phase 3 trial. Lancet 2010; 376(9744):895-902; PMID:20728932; http://dx.doi.org/10.1016/S0140-6736(10)61030-6
8. Wu T, Li SW, Zhang J, Ng MH, Xia NS, Zhao Q. Hepatitis E vaccine development: a 14 year odyssey. Hum Vaccin Immunother 2012; 8(6):823-7; PMID:22699438; http://dx.doi.org/10.4161/hv.20042
9. Xing L, Li TC, Mayazaki N, Simon MN, Wall JS, Moore M, Wang CY, Takeda N, Wakita T, Miyamura T, et al. Structure of hepatitis E virion-sized particle reveals an RNA-dependent viral assembly pathway. J Biol Chem 2010; 285(43):33175-83; PMID:20720013; http://dx.doi.org/10.1074/jbc.M110.106336
10. Li TC, Yoshimatsu K, Yasuda SP, Arikawa J, Koma T, Kataoka M, Ami Y, Suzaki Y, Mai le TQ, Hoa NT, et al. Characterization of self-assembled virus-like particles of rat hepatitis E virus generated by recombinant baculoviruses. J Gen Virol 2011; 92(Pt 12):2830-7; PMID:21865442; http://dx.doi.org/10.1099/vir.0.034835-0
11. Guu TS, Liu Z, Ye Q, Mata DA, Li K, Yin C, Zhang J, Tao YJ. Structure of the hepatitis E virus-like particle suggests mechanisms for virus assembly and receptor binding. Proc Natl Acad Sci U S A 2009; 106 (31):12992-7; PMID:19622744; http://dx.doi.org/10.1073/pnas.0904848106
12. Yamashita T, Mori Y, Miyazaki N, Cheng RH, Yoshimura M, Unno H, Shima R, Moriishi K, Tsukihara T, Li TC, et al. Biological and immunological characteristics of hepatitis E virus-like particles based on the crystal structure. Proc Natl Acad Sci U S A 2009; 106 (31):12986-91; PMID:19620712; http://dx.doi.org/10.1073/pnas.0903699106
13. Li S, Tang X, Seetharaman J, Yang C, Gu Y, Zhang J, Du H, Shih JW, Hew CL, Sivaraman J, et al.

Dimerization of hepatitis E virus capsid protein E2s domain is essential for virus-host interaction. PLoS Pathog 2009; 5(8):e1000537; PMID:19662165; http://dx.doi.org/10.1371/journal.ppat.1000537
14. Tang X, Yang C, Gu Y, Song C, Zhang X, Wang Y, Zhang J, Hew CL, Li S, Xia N, et al. Structural basis for the neutralization and genotype specificity of hepatitis E virus. Proc Natl Acad Sci U S A 2011; 108 (25):10266-71; PMID:21642534; http://dx.doi.org/10.1073/pnas.1101309108
15. Zhang J, Gu Y, Ge SX, Li SW, He ZQ, Huang GY, Zhuang H, Ng MH, Xia NS. Analysis of hepatitis E virus neutralization sites using monoclonal antibodies directed against a virus capsid protein. Vaccine 2005; 23(22):2881-92; PMID:15780737; http://dx.doi.org/10.1016/j.vaccine.2004.11.065
16. Li SW, Zhang J, Li YM, Ou SH, Huang GY, He ZQ, Ge SX, Xian YL, Pang SQ, Ng MH, et al. A bacterially expressed particulate hepatitis E vaccine: antigenicity, immunogenicity and protectivity on primates. Vaccine 2005; 23(22):2893-901; PMID:15780738; http://dx.doi.org/10.1016/j.vaccine.2004.11.064
17. Zhao Q, Li S, Yu H, Xia N, Modis Y. Virus-like particle-based human vaccines: quality assessment based on structural and functional properties. Trends Biotechnol 2013; 31(11):654-63; PMID:24125746; http://dx.doi.org/10.1016/j.tibtech.2013.09.002
18. Grgacic EV, Anderson DA. Virus-like particles: passport to immune recognition. Methods (San Diego, Calif) 2006; 40(1):60-5; PMID:16997714; http://dx.doi.org/10.1016/j.ymeth.2006.07.018
19. Yang C, Pan H, Wei M, Zhang X, Wang N, Gu Y, Du H, Zhang J, Li S, Xia N. Hepatitis E virus capsid protein assembles in 4M urea in the presence of salts. Protein Sci 2013; 22(3):314-26; PMID:23281113; http://dx.doi.org/10.1002/pro.2213
20. Zhang X, Wei M, Pan H, Lin Z, Wang K, Weng Z, Zhu Y, Xin L, Zhang J, Li S, et al. Robust manufacturing and comprehensive characterization of recombinant hepatitis E virus-like particles in Hecolin((R)). Vaccine 2014; 32(32):4039-50; PMID:24892250; http://dx.doi.org/10.1016/j.vaccine.2014.05.064
21. Wu T, Wu XL, Ou SH, Lin CX, Cheng T, Li SW, Ng MH, Zhang J, Xia NS. Difference of T cell and B cell activation in two homologous proteins with similar antigenicity but great distinct immunogenicity. Mol Immunol 2007; 44(12):3261-6; PMID:17408743; http://dx.doi.org/10.1016/j.molimm.2007.01.002
22. Zhang J, Shih JW, Wu T, Li SW, Xia NS. Development of the hepatitis E vaccine: from bench to field. Semin Liver Dis 2013; 33(1):79-88; PMID:23564392; http://dx.doi.org/10.1055/s-0033-1338116
23. Zhao Q, Zhang J, Wu T, Li SW, Ng MH, Xia NS, Shih JW. Antigenic determinants of hepatitis E virus and vaccine-induced immunogenicity and efficacy. J Gastroenterol 2013; 48(2):159-68; PMID:23149436; http://dx.doi.org/10.1007/s00535-012-0701-1
24. Deschuyteneer M, Elouahabi A, Plainchamp D, Plisnier M, Soete D, Corazza Y, Lockman L, Giannini S, Deschamps M. Molecular and structural characterization of the L1 virus-like particles that are used as vaccine antigens in Cervarix, the AS04-adjuvanted HPV-16 and -18 cervical cancer vaccine. Hum Vaccin 2010; 6(5):407-19; PMID:20953154; http://dx.doi.org/10.4161/hv.6.5.11023
25. Mulder AM, Carragher B, Towne V, Meng Y, Wang Y, Dieter L, Potter CS, Washabaugh MW, Sitrin RD, Zhao Q. Toolbox for non-intrusive structural and functional analysis of recombinant VLP based vaccines: a case study with hepatitis B vaccine. PloS one 2012; 7 (4):e33235; PMID:22493667; http://dx.doi.org/10.1371/journal.pone.0033235
26. Wei M, Zhang X, Yu H, Tang ZM, Wang K, Li Z, Zheng Z, Li S, Zhang J, Xia N, et al. Bacteria expressed hepatitis E virus capsid proteins maintain virion-like epitopes. Vaccine 2014; 32(24):2859-65; PMID:24662711; http://dx.doi.org/10.1016/j.vaccine.2014.02.025
27. Zhang Y, Li M, Yang F, Li Y, Zheng Z, Zhang X, Lin Q, Wang Y, Li SW, Xia NS, et al. Comparable quality attributes of hepatitis E vaccine antigen with and without adjuvant adsorption-dissolution treatment, Hum Vaccin Immunothare, 2015, In Press.
28. Towne V, Zhao Q, Brown M, Finnefrock AC. Pairwise antibody footprinting using surface plasmon resonance technology to characterize human papillomavirus type 16 virus-like particles with direct anti-HPV antibody immobilization. J Immunol Methods 2013; 388(1-2):1-7; PMID:23159495; http://dx.doi.org/10.1016/j.jim.2012.11.005
29. Zhao Q, Potter CS, Carragher B, Lander G, Sworen J, Towne V, Abraham D, Duncan P, Washabaugh MW, Sitrin RD. Characterization of virus-like particles in GARDASIL(R) by cryo transmission electron microscopy. Hum Vaccin Immunother 2014; 10(3):734-9; PMID:24299977; http://dx.doi.org/10.4161/hv.27316
30. Shank-Retzlaff M, Wang F, Morley T, Anderson C, Hamm M, Brown M, Rowland K, Pancari G, Zorman J, Lowe R, et al. Correlation between mouse potency and in vitro relative potency for human papillomavirus Type 16 virus-like particles and Gardasil vaccine samples. Hum Vaccin 2005; 1(5):191-7; PMID:17012876; http://dx.doi.org/10.4161/hv.1.5.2126
31. Schofield T. In vitro versus in vivo concordance: a case study of the replacement of an animal potency test with an immunochemical assay. Dev Biol 2002; 111:299-304; PMID:12678253
32. Sitrin RD, Zhao Q, Potter CS, Carragher B, Washabaugh MW. Recombinant protein based viral vaccines, In Vaccine Analysis: Strategies, Principles and Control. Brian Nunnally et al. (eds.) Springer Verlag Berkub Heidelberg, 2015. Chapter 3, Recombinant Virus-like Particle Protein Vaccines; p.81-112 (DOI: 10.1007/978-3-662-45024-6-3)
33. Shank-Retzlaff ML, Zhao Q, Anderson C, Hamm M, High K, Nguyen M, Wang F, Wang N, Wang B, Wang Y, et al. Evaluation of the thermal stability of Gardasil. Hum Vaccin 2006; 2(4):147-54; PMID:17012891; http://dx.doi.org/10.4161/hv.2.4.2989
34. Zhao Q, Wang Y, Abraham D, Towne V, Kennedy R, Sitrin RD. Real time monitoring of antigenicity development of HBsAg virus-like particles (VLPs) during heat- and redox-treatment. Biochem Biophys Res Commun 2011; 408(3):447-53; PMID:21527246; http://dx.doi.org/10.1016/j.bbrc.2011.04.048
35. Zhao Q, Towne V, Brown M, Wang Y, Abraham D, Oswald CB, Gimenez JA, Washabaugh MW, Kennedy R, Sitrin RD. In-depth process understanding of RECOMBIVAX HB(R) maturation and potential epitope improvements with redox treatment: multifaceted biochemical and immunochemical characterization. Vaccine 2011; 29(45):7936-41; PMID:21871939; http://dx.doi.org/10.1016/j.vaccine.2011.08.070
36. Zhao Q, Modis Y, High K, Towne V, Meng Y, Wang Y, Alexandroff J, Brown M, Carragher B, Potter CS, et al. Disassembly and reassembly of human papillomavirus virus-like particles produces more virion-like antibody reactivity. Virol J 2012; 9:52; PMID:22356831; http://dx.doi.org/10.1186/1743-422X-9-52
37. Zhang J, Zhang X-F, Huang S-J, Wu T, Hu Y-M, Wang Z-Z, et al. Long-term efficacy of a hepatitis E vaccine. N Engl J Med 2015;372(10):914-22; PMID: 25738667; http://dx.doi.org/10.1056/NEJMoa1406011.
38. Zhang J, Zhang XF, Zhou C, Wang ZZ, Huang SJ, Yao X, Liang ZL, Wu T, Li JX, Yan Q, et al. Protection against hepatitis E virus infection by naturally acquired and vaccine-induced immunity. Clin Microbiol Infect 2014; 20(6):O397-405; PMID:24118636; http://dx.doi.org/10.1111/1469-0691.12419
39. Wu T, Huang SJ, Zhu FC, Zhang XF, Ai X, Yan Q, Wang ZZ, Yang CL, Jiang HM, Liu XH, et al. Immunogenicity and safety of hepatitis E vaccine in healthy hepatitis B surface antigen positive adults. Hum Vaccin Immunother 2013; 9(11):2474-9; PMID:23887167; http://dx.doi.org/10.4161/hv.25814
40. WHO Weekly epidemiological record, 2014, 89, 321-336.
41. Wu T, Zhu FC, Huang SJ, Zhang XF, Wang ZZ, Zhang J, Xia NS. Safety of the hepatitis E vaccine for pregnant women: A preliminary analysis. Hepatology 2012; 55(6):2038; PMID:22161542; http://dx.doi.org/10.1002/hep.25522
42. Nelson KE, Shih JW, Zhang J, Zhao Q, Xia N, Ticehurst J, et al. Hepatitis E vaccine to prevent morbidity and mortality during epidemics. Open Forum Infect Dis 2014:1(3): ofu098; PMID:25714510; PMID: 25734166; http://dx.doi.org/10.1093/ofid/ofu098

EBioMedicine 51 (2020) 102584

Contents lists available at ScienceDirect

EBioMedicine

journal homepage: www.elsevier.com/locate/ebiom

Research paper

Genetic risk for dengue hemorrhagic fever and dengue fever in multiple ancestries

Guillaume Pare[a,b], Binod Neupane[a], Sasha Eskandarian[a], Eva Harris[d], Scott Halstead[e], Lionel Gresh[f], Guillermina Kuan[g], Angel Balmaseda[f,h], Luis Villar[i], Elsa Rojas[j], Jorge E. Osorio[k], Dang Duc Anh[l], Aruna Dharshan De Silva[m], Sunil Premawansa[n], Gayani Premawansa[m], Ananda Wijewickrama[m], Ivette Lorenzana[o], Leda Parham[o], Cynthia Rodriguez[o], Ildefonso Fernandez-Salas[p], Rosa Sanchez-Casas[p], Esteban E. Diaz-Gonzalez[p], Khin Saw Aye[q], Win Lai May[q], Min Thein[q], Filemon Bucardo[r], Yaoska Reyes[r], Patricia Blandon[r], Kenji Hirayama[s], Lan Weiss[s,t], Pardeep Singh[a], Jennifer Newton[a], Mark Loeb[a,b,c,*], for the Dengue Population Genetics Program

[a] *Department of Pathology and Molecular Medicine, McMaster University, Ontario L8N 3Z5, Canada*
[b] *Department of Health Research, Methods, Evidence, and Impact, Canada*
[c] *Institute for Infectious Diseases Research, McMaster University Hamilton, Canada*
[d] *Division of Infectious Diseases and Vaccinology, School of Public Health, University of California, Berkeley, CA, United States*
[e] *Department of Preventive Medicine and Biometrics, Uniformed University of the Health Sciences, Bethesda, MD, United States*
[f] *Sustainable Sciences Institute, Managua, Nicaragua*
[g] *Health Center Sócrates Flores Vivas, Ministry of Health, Managua, Nicaragua*
[h] *Laboratorio Nacional de Virología, Centro Nacional de Diagnóstico y Referencia, Ministry of Health, Managua, Nicaragua*
[i] *Clinical Epidemiology Unit, Universidad Industrial de Santander, Bucaramanga, Colombia*
[j] *Centro de Atención y Diagnóstico de Enfermedades Infecciosas, Bucaramanga, Colombia*
[k] *University of Wisconsin, Wisconsin, United States*
[l] *National Institute of Hygiene and Epidemiology, Hanoi, Vietnam*
[m] *Genetech Research Institute, Sri Lanka*
[n] *Department of Zoology and Environmental Sciences, University of Colombo, Sri Lanka*
[o] *Department of National Autonomous University of Honduras, Tegucigalpa, Honduras*
[p] *Universidad Autonoma de Nuevo Leon, Mexico*
[q] *Medical Research, Ministry of Health, Myanmar*
[r] *The Faculty of Medical Sciences at the National Autonomous University of León, Nicaragua*
[s] *Department of Immunogenetics, Institute of Tropical Medicine, Nagasaki University, Nagaski, Japan*
[t] *Department of Immunology and Microbiology, Pasteur Institute, Ho Chi Minh City, Vietnam*

ARTICLE INFO

Article History:
Received 8 August 2019
Revised 28 October 2019
Accepted 26 November 2019
Available online xxx

Keywords:
Dengue
Genetics
Risk
GWAS
Ancestry

ABSTRACT

Background: Genetic risk factors for dengue hemorrhagic fever/dengue shock syndrome (DHF/DSS) and dengue fever (DF) are limited, in particular there are sparse data on genetic risk across diverse populations.
Methods: We conducted a genome-wide association study (GWAS) in a derivation and validation sample of 7,460 participants of Latin American, South Asian, and South East Asian ancestries. We then developed a weighted polygenic risk score (PRS) for each participant in each of the validation cohorts of the three ancestries to predict the risk of DHF/DSS compared to DF, DHF/DSS compared to controls, and, DF compared to controls.
Findings: The risk of DHF/DSS was significantly increased, odds ratio [OR] 1.84 (95%CI 1.47 to 2.31) (195 SNPs), compared to DF, fourth PRS quartile versus first quartile, in the validation cohort. The risk of DHF/DSS compared to controls was increased (OR=3.94; 95% CI 2.84 to 5.45) (278 SNPs), as was the risk of DF compared to controls (OR=1.97; 95%CI 1.63 to 2.39) (251 SNPs). Risk increased in a dose-dependent manner with increase in quartiles of PRS across comparisons. Significant associations persisted for PRS built within ancestries and applied to the same or different ancestries as well as for PRS built for one outcome (DHF/DSS or DF) and applied to the other.

---

* Corresponding author at: Department of Pathology and Molecular Medicine, McMaster University Hamilton, Ontario L8N 3Z5, Canada.
E-mail address: loebm@mcmaster.ca (M. Loeb).

https://doi.org/10.1016/j.ebiom.2019.11.045
2352-3964/© 2019 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license. (http://creativecommons.org/licenses/by-nc-nd/4.0/)

Interpretation: There is a strong genetic effect that predisposes to risk of DHF/DSS and DF. The genetic risk for DHF/DSS is higher than that for DF when compared to controls, and this effect persists across multiple ancestries.

© 2019 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license. (http://creativecommons.org/licenses/by-nc-nd/4.0/)

Research in Context

Evidence before this study

We conducted a literature review on MEDLINE in January 2013 before conducting this study on genetic determinants of dengue and did not find a systematic review. We found narrative reviews and individual candidate gene studies. Over the course of the study and after completion we updated our search (last search, November 5, 2018). We identified candidate gene studies but only one GWAS study in a Vietnamese population, replicated in a Vietnamese and in a Thai population, demonstrating as association between variants in major histocompatibility complex (MHC) class I polypeptide-related sequence B (*MICB)* and phospholipase C, epsilon 1 (*PLCE1)* and DSS. We did not identify any studies addressing polygenic risk for dengue.

Added value of this study

This study demonstrates a strong polygenic risk for dengue that applies across different ancestries. The polygenic risk derived from one syndrome predicted risk in the validation cohort of another syndrome.

Implications of all the available evidence

In certain populations, specific genetic variants are associated with severe complications of dengue. However, there is genetic susceptibility that is universal, that is, the same alleles increase or reduce susceptibility in diverse regions of the world. This is suggestive of a more general immune mechanism rather than adaptation to DENV-specific strain effects, which would be expected given the large geographical distribution of participants. Moreover, the effect of a similar risk when a polygenic risk score is derived from one syndrome and applied to the other, suggests a common pathological mechanism.

1. Introduction

Infection with dengue virus (DENV), a mosquito-borne flavivirus infection, is of major global public health relevance [1]. Dengue fever (DF), or “break-bone fever”, is characterized by fever, headache, retro-ocular, and joint pain, rash, and, lymphadenopathy [2]. In less than 2% of those infected for a second time, the disease may progress to dengue hemorrhagic fever characterized by thrombocytopenia and vascular leakage and dengue shock syndrome (DSS) (DHF with evidence of systemic hypoperfusion) [3–8]. A secondary heterologous infection, that is infection with a DENV serotype different from that of prior infection, can increase risk of DHF/DSS through antibody-dependent enhancement [9,10]. However, given that only a small proportion of secondary infections result in severe disease, underlying genetic predisposition is likely [11].

The vast majority of genetic studies for dengue have been candidate gene studies, where a number of genetic variants, including single nucleotide polymorphisms and HLA polymorphisms, have been implicated as genetic risk factors [12–19]. Existing data have been conflicting [20]. There has only been one genome-wide association study (GWAS) conducted for dengue, which showed an association between variants in major histocompatibility complex (MHC) class I polypeptide-related sequence B (*MICB)* and phospholipase C, epsilon 1 (*PLCE1)* and DSS [21]. However, this study was conducted in a single ancestry population (Vietnamese). Although these variants were replicated when tested alone in another Vietnamese and a Thai population [22,23], less is known about the extent to which genetic variants that predispose for dengue act across ancestries. We report here a GWAS in a sample set with multiple ancestries, which allowed us to test the hypothesis, using polygenic risk scores, that genetic variants across the genome predispose to DHF/DSS and DF.

2. Methods and methods

2.1. Study participants

The Dengue Population Genetics Program (DPGP) was a genetic epidemiologic study where DNA was obtained from participants of three major ancestries from seven countries: Latin American from Honduras, Mexico, Nicaragua, Colombia; South Asian from Sri Lanka; and Southeast Asian from Vietnam and Myanmar between March 2003 and December 2013. Participants were categorized into one of the following groups: DF, DHF/DSS, and controls (definitions are shown in Table S1 in the Supplementary Appendix). Clinical phenotype data was obtained from community cohort studies and hospitalized patients with DHF/DSS and DF. Infection was confirmed by serology, DENV RNA demonstrated by RT-PCR, DENV NS1 antigen detection, or, viral isolation [24, 25]. Controls were defined as participants who had evidence of DENV infection but no evidence of DHF/DSS or DF (they consisted of two groups, a group of 513 that seroconverted during the study and 1706 IgG antibody positive participants). They were derived by flow cytometry-based assays or ELISA in cross-sectional samples of participants who had no symptoms of dengue at the time of testing and had no history of hospitalization for dengue, or annual healthy samples by neutralization assay (using plaque reduction neutralization assay or flow cytometry-based assays) for participants in cohort studies [26, 27]. The analysis included participants with primary and secondary DENV infections. A summary of laboratory testing for infection is provided in Table S2 in the Supplementary Appendix.

2.2. Genotyping

We conducted the study using a two-stage design. In the first stage, we used the Illumina OmniExpress BeadChip array of 958,178 single nucleotide polymorphisms (SNPs) with genome-wide coverage to identify genetic associations in three separate comparisons: DHF/DSS versus DF, DHF/DSS versus controls, and, DF versus controls using a derivation cohort of 2498 subjects. We included genetic variants that had previously been associated with DHF/DSS or DF (e.g., DC-SIGN, vitamin D receptor, Fc gamma receptor II, TNF alpha, IL-10, HLA-A, HLA-B, TAP 1, Tap 2, CTLA-4, *MICB* and, *PLCE1)* in the derivation analyses [11, 21]. In the second stage, SNPs with p-values $\leq$0.0001 for at least one of three comparisons in the analysis of the entire derivation cohort were pruned at the linkage disequilibrium cut-off of r^2=0.20 using a 500 kb window. These, along with the genetic variants previously associated DHF/DSS or DF, were genotyped in a validation (replication) cohort of 6879 subjects using the Illumina GoldenGate BeadXpress platform using a total of 1536 SNPs.

2.3. Quality control

In the overall derivation cohort, samples with <99% call rates and sex mismatch were removed, as were the SNPs with <99% call rate, <1% minor allele frequency (MAF) or Hardy-Weinberg p-value <0.000001 in controls. For the ancestry-specific derivation cohort, we further removed SNPs with MAF <5% because the sample size was limited. In the overall validation cohort, we removed samples with <95% call rates, SNPs with <95% call rate, or SNPs with <1% MAF in order to avoid excluding SNPs that were narrowly < 5%. The quality controls in derivation and validation cohorts were performed using PLINK (version 1.09).

A total of 9377 DNA samples was received for analysis, of which for the derivation cohort 999 were from participants with DHF/DSS, 999 DF, and, 500 were from controls. For the validation cohort, there were 1715 samples from participants with DHF, 2731 DF, and, 2433 controls. Following quality control steps, there were a total of 2248 DNA samples from the derivation phase and 5212 samples from the validation phase with the final dataset consisting of 7460 samples. The characteristics of participants in this final dataset are summarized in Table 1. Of the total of 958,178 SNPs genotyped in the derivation phase, 662,390 met quality control and the MAF criteria, as did 1165 of the 1536 SNPs tagged for the validation phase.

2.4. Association testing and replication

We calculated heritability using Genome Wide Complex Trait Analysis [28]. We used the Grammar-Gamma method to test for the associations of SNPs with DHF/DSS versus DF, DF versus controls, and, DHF/DSS versus controls in the entire derivation cohort (all ancestries), adjusted for age and sex, and population structure was taken into account by the kinship coefficient matrix of the variance component model [29]. The Grammar-Gamma analysis was also performed separately for each derivation sample of Latin American, South Asian, and, Southeast Asian ancestries, where only SNPs with allele frequency ≥5% were analyzed and a heritability coefficient (h^2) of 0.3 was assumed (rather than estimated from the data) in a sample to ease model convergence as ancestry-specific sample sizes were small. We also checked the data for outliers from ancestries other than Latin American, South Asian, and Southeast Asian. We used a significance level of $p<5 \times 10^{-8}$ to account for multiple testing. In the replication cohort, a separate logistic regression model was fitted adjusting for age, sex and ancestry for each SNP for each comparison. In order to exclude the possibility of population stratification, we did the analyses adjusted for principal components.

Table 1
Distribution of participants by age, sex, and ethnicity.

	Ethnicity			
Characteristics	Latin America	South Asia	South East Asia	Total
Derivation cohort				
Control	210	44	151	405
Female sex,%	61.4	47.7	39.1	51.6
Mean age (SD), yr	23.2 (18.5)	35.4 (14.3)	27.0 (21.9)	26.0 (19.8)
Cases				
DF	498	119	308	925
Female sex,%	57.8	35.3	44.8	50.6
Mean age (SD), yr	21.7 (16.9)	29.5 (15.4)	12.6 (11.0)	19.7 (16.0)
DHF/DSS	457	92	369	918
Female sex,%	49.5	50.0	50.9	50.1
Mean age (SD), yr	22.1 (16.7)	27.3 (12.9)	11.9 (8.0)	18.5 (14.6)
Replication cohort				
Control	1449	127	238	1814
Female sex,%	60.9	48.0	55.0	59.3
Mean age (SD), yr	22.1 (18.0)	37.2 (16.1)	31.5 (19.7)	24.4 (18.7)
Cases				
DF	1386	354	361	2101
Female sex,%	55.8	37.6	51.0	51.9
Mean age (SD), yr	22.3 (18.2)	29.4 (12.8)	7.9 (8.4)	21.0 (17.4)
DHF/DSS	295	233	769	1297
Female sex,%	56.9	42.1	48.0	49.0
Mean age (SD), yr	25.6 (16.1)	27.0 (12.0)	8.9 (7.4)	16.0 (13.8)

Abbreviation: SD, standard deviation; DF, Dengue fever; DHF, Dengue hemorrhagic fever; DSS, Dengue Shock Syndrome.

2.5. Polygenic risk score analysis

To test for the possibility of polygenic risk for DENV infection outcome (DHF/DSS versus DF, DHF/DSS versus control, and DF versus control), we developed a weighted polygenic risk score (PRS) for each participant in each of the validation cohorts of three ancestries (Latin American, South Asian, Southeast Asian) separately. For each participant, the PRS was the sum of the product of log-odds ratios of a set of SNPs passing a p-value threshold in a derivation cohort and the number of effect alleles at the corresponding SNPs (see Methods section in the Supplementary Appendix). Polygenic risk scores were constructed using all three ancestries (the primary PRS analysis) as well as single ancestries for sensitivity analyses. We also conducted the analysis using five principal components as covariates. If the allele at a SNP for a participant was missing, the number of effect alleles was imputed using the average number of effect alleles in all participants within the same ancestry with the same outcome. Nine different p-value thresholds ranging from 0.1 to 0.00001 were used for selecting SNPs to compute PRSs. All SNPs had a p-value < 0.001 for at least one analysis but could have had larger p-values for other comparisons.

2.6. Overall effect

To test the association of PRS with each of three comparisons (DHF/DSS versus DF, DHF/DSS versus controls and DF versus controls), first a logistic regression model was fitted by comparing the second to fourth quartile PRS to the first (referent), adjusting for age and sex in each validation ancestry. Based on plots of the p-value threshold and effect size for selecting SNPs for the PRS, which remained similar from a range of 0.1 to 0.001, we selected a p-value of 0.01 for the PRS (Figure S1 in the Supplementary Appendix). To assess the association of PRS in the overall validation cohort (overall effect), the results (the log-odds ratios and corresponding standard errors) from all three ancestries were then pooled using the fixed-effect model of meta-analysis. We also conducted a similar set of analyses by demonstrating the effect using a standardized PRS (i.e., PRS divided by its standard deviation, SD) adjusting for age and sex in each validation ancestry.

2.7. Secondary analyses

We conducted secondary analyses using derivation and validation cohorts of different ancestries. We examined 1) the effect of a PRS derived from all three ancestries and applied to a validation cohort of a single ancestry, 2) the effect of a PRS from the derivation cohort of a single ancestry applied to the validation cohort of the same ancestry, 3) the association of PRS from one derivation ancestry applied to a validation cohort of two different ancestries, 4) the effect of PRS derived from one single ancestry applied to a validation cohort of all ancestries, 5) a PRS from the derivation cohort of either DHF/DSS versus control or DF versus control then applied to the validation cohort of the other syndrome comparison. We examined these using a standardized PRS with the OR expressed per change in SD or by comparing quartiles.

The Grammar-Gamma analyses were performed using the R package GenABEL (version 1.8–0). Statistical Gene score analyses were performed using R software, version 3.2.4.

2.8. Role of the funding source

The study sponsor played no role in study design; in the collection, analysis, and interpretation of data; in the writing of the report; nor in the decision to submit the paper for publication.

2.9. Ethics statement

This study was approved by the McMaster research ethics board and that of all participating institutions. Participants provided informed consent for the analyses done in this study.

3. Results

3.1. Heritability estimates

By testing for chip-wide heritability, that is, the portion of the phenotypic variation that can be explained by genotyped genetic markers [30], overall and by ancestry, we found significant values for DHF/DSS compared to DF for Latin Americans and Southeast Asians (0.18 and 0.21), for DHF/DSS compared to controls for Latin Americans, South Asians, and Southeast Asians (0.35, 0.21, 0.18 respectively), and, for DF compared to controls for Latin Americans and Southeast Asians (0.20 and 0.56 respectively), all p-values <0.05. The establishment of such heritability supported testing for genetic risk of variants.

3.2. Association testing

In the analysis of the entire derivation sample, none of the SNPs reached the significance level of $p<5 \times 10^{-8}$ for any comparison (Figure S2, Table S3 in the Supplementary Appendix). For the DHF/DSS versus DF comparison, the top associated SNP in the meta-analysis of derivation and replication phases was rs6675033 with p-value of 2.40×10^{-7} near the transcriptional adapter 1 (*TADA1*) gene on Chromosome 1 (Table S3), where the association was observed only in the derivation but not in replication sample. Genetic variants tested that had previously been found associated with DHF/DSS or DF (DC-SIGN, vitamin D receptor, Fc gamma receptor II,TNF alpha, IL-10, HLA-A, HLA-B, TAP 1, Tap 2, CTLA-4, *MICB* and *PLCE1*) did not reach the pre-specified significance in the derivation analysis and were not replicated. Including ancestry as a co-variate did not change the results. We also conducted the genome-wide analysis separately in a meta-analysis setting, including using ancestry specific principal components, and the findings remained similar. The presence of significant heritability yet no genome-wide significance suggested a polygenic model. On this basis, we tested a PRS.

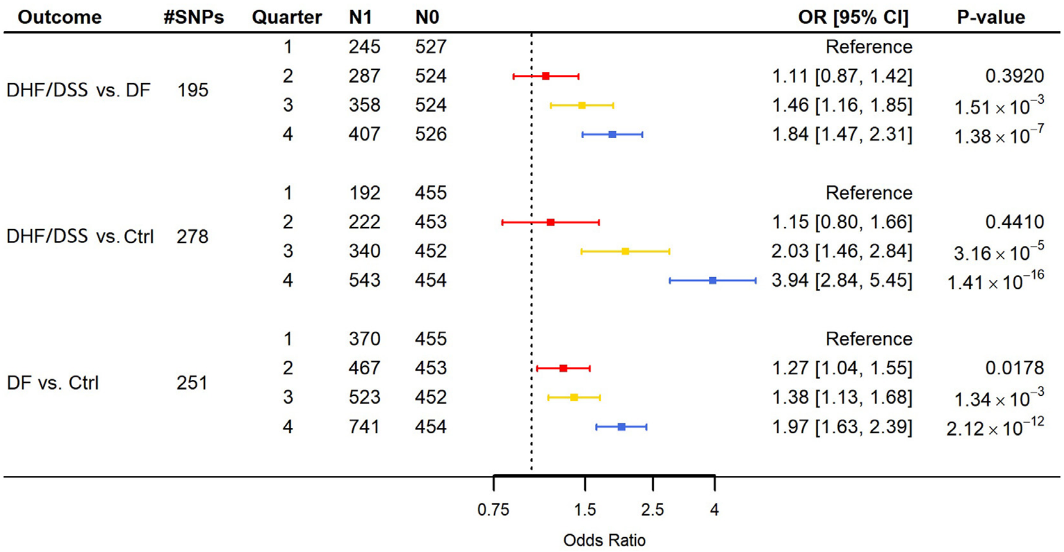
3.3. Polygenic risk score

The risk of DHF/DSS was significantly increased, odds ratio [OR] 1.84 (95%CI 1.47 to 2.31), p-value 1.38×10^{-7} compared to DF, when the fourth-quartile PRS was compared to the first-quartile PRS (referent) in the validation cohort (Fig. 1). The risk of DHF/DSS compared to controls was increased, OR 3.94 (95% CI 2.84 to 5.45), p-value 1.4×10^{-16}, for the fourth-quartile PRS to reference comparison. Similarly, the risk for DF increased compared to controls, OR 1.97 (95%CI 1.63 to 2.39), p-value 2.12×10^{-12}, fourth-quartile PRS compared to reference. The effect size increased in a dose-dependent manner as the PRS increased, with a highly significant effect for third quartile versus referent comparison for all three outcomes (Fig. 1). Although ORs >1 for the PRS in the second quartile compared to the first was observed for all three outcomes, they were only significant for DF versus controls.

Using a standardized PRS, the results were similar (Table 2, Fig. 2). The effect of PRS using a p-value threshold of 0.01 resulted in the following effects: OR 1.31 (1.20 to 1.42) per SD for DHF/DSS versus DF, p-value 1.95×10^{-11}; OR 1.89 (1.69 to 2.11) per SD for DHF/DSS versus control, p-value $1.49 \times 10^{-33;}$ OR 1.29 (1.21 to 1.38) per SD, DF versus control, p-value 5.80 x 10^{-13}. Using a different p-value threshold of 0.0005 to derive the PRS resulted in similar effect sizes: OR 1.29 (1.19 to 1.40) for DHF/DSS versus DF, p-value 2.50×10^{-10}; OR 1.76 (1.57 to 1.96) for DHF/DSS versus control, p-value 3.87×10^{-27}; OR 1.22 (1.14 to 1.31), DF versus control, p-value 1.28×10^{-8}. The effect size of all p-value thresholds is shown in Figure S1. Our results did not change when we included principal components as covariates.

3.4. Secondary analyses

For each of the three DENV infection outcome comparisons (DHF/DSS versus DF, DHF/DSS versus control, DF versus control), the results of our five secondary analyses were similar in direction and



Outcome	#SNPs	Quarter	N1	N0	OR [95% CI]	P-value
DHF/DSS vs. DF	195	1	245	527	Reference	
		2	287	524	1.11 [0.87, 1.42]	0.3920
		3	358	524	1.46 [1.16, 1.85]	1.51×10^{-3}
		4	407	526	1.84 [1.47, 2.31]	1.38×10^{-7}
DHF/DSS vs. Ctrl	278	1	192	455	Reference	
		2	222	453	1.15 [0.80, 1.66]	0.4410
		3	340	452	2.03 [1.46, 2.84]	3.16×10^{-5}
		4	543	454	3.94 [2.84, 5.45]	1.41×10^{-16}
DF vs. Ctrl	251	1	370	455	Reference	
		2	467	453	1.27 [1.04, 1.55]	0.0178
		3	523	452	1.38 [1.13, 1.68]	1.34×10^{-3}
		4	741	454	1.97 [1.63, 2.39]	2.12×10^{-12}

Fig. 1. Association of polygenic risk scores with dengue virus infection outcomes. Notations/abbreviations: N1 and N0, total number of cases and comparisons in the validation cohort; P, p-value; OR, odds ratio; CI, confidence interval. *Both derivation and validation cohorts were entire samples of all three ancestries (Latin American, South Asian and Southeast Asian). For each outcome comparison, the estimate of odds ratio of per standard deviation increase in polygenic risk score in the overall validation cohort was obtained by pooling the corresponding log-OR estimates across three ancestries using a fixed-effect meta-analysis model.

Table 2
Pooled association of PRS derived from overall derivation cohort in overall validation cohort.

Outcome*	Total	Cases	Comparators	Threshold	#SNPs	Polygenic risk score† OR [95%CI]	Polygenic risk score† p-value
DF vs. Ctrl	3915	2101	1814	0.1	387	1.33[1.24, 1.43]	2.20e-16
				0.05	332	1.33[1.25, 1.43]	1.50e-16
				0.01	251	1.29[1.21, 1.38]	1.99e-13
				0.005	222	1.27[1.19, 1.36]	5.17e-12
				0.001	179	1.24[1.16, 1.32]	8.11e-10
				0.0005	155	1.22[1.14, 1.31]	6.71e-09
				0.0001	24	1.13[1.05, 1.20]	4.63e-04
				0.00005	4	1.00[0.93, 1.07]	9.29e-01
				0.00001	2	0.99[0.93, 1.06]	7.61e-01
DHF/DSS vs.Ctrl	3111	1297	1814	0.1	434	1.87[1.68, 2.09]	1.52e-28
				0.05	374	1.87[1.67, 2.09]	2.31e-28
				0.01	278	1.89[1.69, 2.11]	3.09e-29
				0.005	256	1.88[1.68, 2.10]	4.52e-29
				0.001	199	1.75[1.56, 1.95]	2.98e-23
				0.0005	176	1.76[1.57, 1.96]	9.71e-24
				0.0001	33	1.43[1.29, 1.59]	9.11e-12
				0.00005	13	1.21[1.10, 1.34]	1.30e-04
				0.00001	1	1.09[0.97, 1.23]	1.47e-01
DHF/DSS vs.DF	3398	1297	2101	0.1	317	1.35[1.24, 1.46]	5.63e-13
				0.05	265	1.34[1.23, 1.45]	3.07e-12
				0.01	195	1.31[1.20, 1.42]	1.16e-10
				0.005	183	1.29[1.19, 1.40]	6.78e-10
				0.001	167	1.29[1.19, 1.40]	7.62e-10
				0.0005	136	1.29[1.19, 1.40]	8.65e-10
				0.0001	15	1.08[1.00, 1.17]	4.76e-02
				0.00005	7	1.05[0.97, 1.14]	2.25e-01
				0.00001	2	1.07[0.99, 1.16]	1.10e-01

Abbreviations: DF, dengue fever, DHF/DSS, dengue hemorrhagic fever; OR, odds ratio; CI, confidence interval.

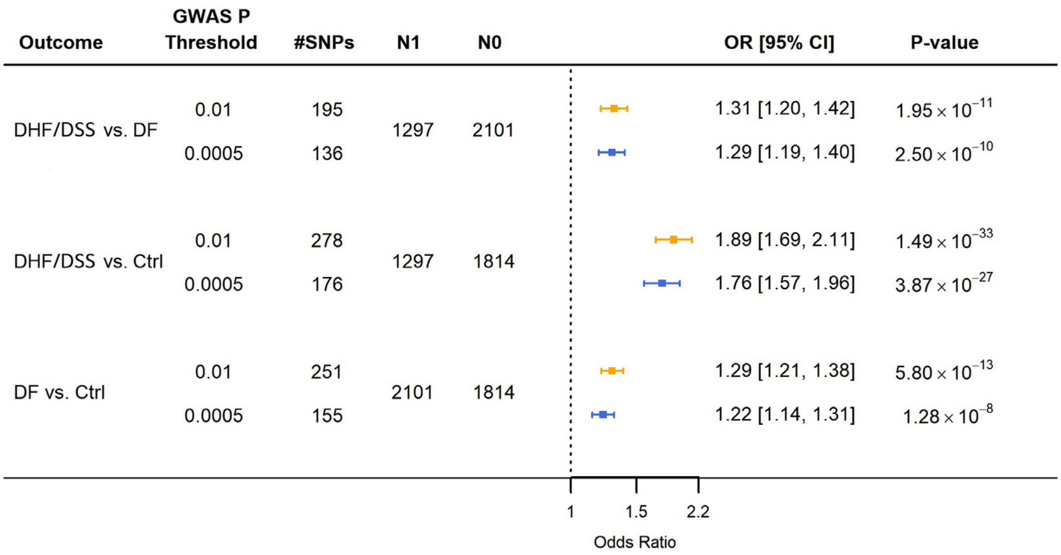
* Outcomes for the derivation and validation cohorts were the same. Figures are from replication cohort.

† Age- and sex-adjusted log(OR) estimates per standard deviation (SD) increase in polygenic risk score across Latin American, South Asian, and Southeast Asian validation cohorts were meta-analyzed using fixed-effect model.

magnitude to our overall effects (Tables S4 -S9, Figures S3-S6). When we derived the PRS from the overall derivation cohort for DF versus control and applied it to a validation cohort for DHF versus control, the effect size was similar to our previous estimates, OR 1.60 (95%CI 1.45 to 1.77), p-value 2×10^{-19} (Table S9). Conversely, when the PRS was derived from DHF/DSS versus control and applied to a validation cohort of DF versus control, the effect was similar, OR 1.37 (95%CI 1.28 to 1.46), p-value 8.45×10^{-19} .

4. Discussion

We found a strong genetic risk, derived from PRS from a multi-ancestry population, when DHF/DSS was compared to DF, DHF/DSS compared to controls, and, DF compared to controls. Moreover, this polygenic risk had a dose effect, with the risk increasing with an increasing number of risk alleles in the PRS. Sensitivity analyses demonstrated that PRS derived from single ancestries and applied to a



Outcome	GWAS P Threshold	#SNPs	N1	N0	OR [95% CI]	P-value
DHF/DSS vs. DF	0.01	195	1297	2101	1.31 [1.20, 1.42]	1.95×10^{-11}
	0.0005	136			1.29 [1.19, 1.40]	2.50×10^{-10}
DHF/DSS vs. Ctrl	0.01	278	1297	1814	1.89 [1.69, 2.11]	1.49×10^{-33}
	0.0005	176			1.76 [1.57, 1.96]	3.87×10^{-27}
DF vs. Ctrl	0.01	251	2101	1814	1.29 [1.21, 1.38]	5.80×10^{-13}
	0.0005	155			1.22 [1.14, 1.31]	1.28×10^{-8}

Fig. 2. Forest plot of the associations of polygenic risk scores* quartiles. Notations/abbreviations: N1 and N0, total number of cases and comparisons in overall validation cohort; P, p-value; OR, odds ratio; CI, confidence interval. *Both derivation and validation cohorts were entire samples of all three ancestries of Latin American, South Asian and Southeast Asian. For an outcome, the estimate of odds ratio of each polygenic risk quarter versus the first quarter was obtained by pooling the corresponding log-OR estimates at p threshold of 0.01 across three ancestries using a fixed-effect meta-analysis model.

global validation cohort remained significant predictors of risk, as did multi-ancestry PRS when applied to single ancestries, and single ancestries PRS when applied to the same ancestry in the validation cohort. We also found that PRS derived from one syndrome predicted risk in the validation cohort of another syndrome.

The effect sizes we observed, such as almost a four-fold risk of DHF/DSS compared to controls, are among the highest risk factors reported for dengue [3]. Studies that have tested variants shown to be significant in a GWAS study reported effect sizes that were substantially lower (< 1.5) [21–23]. Even when examining other risk factors for severe dengue, including age, sex, and, nutritional status, our PRS is the strongest risk factor [31]. The large size of this effect is compatible with a polygenic effect for all three DENV infection outcome comparisons, i.e. DHF/DSS versus DF, DHF/DSS versus controls, DF versus controls. In the overall validation cohort of all three ancestries, the evidence of polygenic risk was stronger for DHF/DSS versus control compared to DF versus control when derived from overall or ancestry-specific GWAS cohorts, suggesting that not only the risk is polygenic but also that the risk of more severe complications increases with higher burden of associated genetic variants.

Our secondary analyses, where we derived the polygenic risk scores from one ancestry population or degree of severity of DENV infection to validate in another ancestry or another severity level, indicate that the association of polygenic risk score is indeed robust. Given the polygenic origin of these traits with large numbers of SNPs with very small effects, it is not surprising that our GWAS findings were not replicated, and this may help explain inconsistencies in the results of previous genetic association studies [11]. Importantly, these secondary analyses demonstrate that the genetic susceptibility we observed is universal, that is, the same alleles increase or reduce susceptibility in diverse regions of the world. This is suggestive of a more general immune mechanism rather than adaptation to DENV-specific strain effects, which would be expected given the large geographical distribution of participants. Moreover, the effect of a similar risk when PRS was derived from one syndrome (DHF/DSS or DF) and applied to the other, suggests a common pathological mechanism.

Strengths of this study are the diverse ancestry of participants and the consistency of the findings. We acknowledge that use of participants that had no symptoms at the time of testing and no history of hospitalization as controls may have led to misclassification persons with DF as controls. However, this would have been mitigated by our use of controls with inapparent infection and serological evidence of infection. Any bias would have been towards the null, but the fact that we found an effect suggests that it in fact may have been underestimated. Furthermore, the effect of misclassification of controls in genetic association studies is minimal [32]. Similarly, any variance across community cohort studies for admission trends would be mitigated by using strict phenotype definitions adjudicated centrally. If thresholds for admissions varied for ascertainment of cases, any significant effect that we reported would be more likely to be underestimated in situations where thresholds for admission of DHF or DSS were lower. Moreover, since the main effects are polygenetic, it is very unlikely that admission to hospital would be associated with any such effect. We acknowledge that the power to detect variants significant at the genome-wide level may have been limited. However, the high degree of heritability and robust estimates of risk using the PRS argue strongly in favor of a polygenic model. We also acknowledge, despite the relatively high effect size, that the odds ratios are however modest for public health application.

In summary, our data provide strong evidence that multiple genetic variants confer clinically important risk for both DHF/DSS and DF. To our knowledge, this is the first report to demonstrate such polygenic risk for an infectious disease. Our data offer insight into how genetic variants predispose to dengue as well as the utility of using PRS in the study of infectious diseases.

Declaration of competing interest

Dr. Pare reports other from Amgen, other from Sanofi, outside the submitted work; the other authors have no conflicts of interest.

Acknowledgments

We thank Laura Childerhose from McMaster University. We thank Andrea Núñez and William Avilés for overseeing quality and data management of the studies based in Nicaragua, respectively. We thank the study teams of the Nicaraguan Pediatric Dengue Cohort Study and the Pediatric Dengue Hospital-based Study at the Centro de Salud Sócrates Flores Vivas (Nicaraguan Ministry of Health), the Hospital Infantil Manuel de Jesús Rivera (Nicaraguan Ministry of Health), the Laboratorio Nacional de Virología at the Centro Nacional de Diagnóstico y Referencia (Nicaraguan Ministry of Health), and the Sustainable Sciences Institute, as well as the study participants and their families.

Funding sources

Funding for the study was provided by the National Institute of Health (HHSN272201000026C). The funders had no role in study design, data collection, data analysis, interpretation, writing of the report.

Supplementary materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.ebiom.2019.11.045.

References

[1] Simmons CP, Farrar JJ, Nguyen vV, Wills B. Dengue. N Engl J Med 2012;366:1423–32.

[2] World Health Organization (2016). Dengue and severe dengue. Available online at: http://www.who.int/mediacentre/factsheets/fs117/en/(Accessed April 22, 2018).

[3] Sangkawibha N, Rojanasuphot S, Ahandrik S, et al. Risk factors in dengue shock syndrome: a prospective epidemiologic study in Rayong, Thailand. I. The 1980 outbreak. Am J Epidemiol 1984 Nov;120(5):653–69.

[4] Halstead SB. Antibody, macrophages, dengue virus infection, shock, and hemorrhage: a pathogenetic cascade. Rev Infect Dis 1989;11(Suppl 4:):S830–S9.

[5] Halstead SB, Porterfield JS, O'Rourke EJ. Enhancement of dengue virus infection in monocytes by flavivirus antisera. Am J Trop Med Hyg 1980;29:638–42.

[6] Kurane I, Ennis FA. Immunity and immunopathology in dengue virus infections. Semin Immunol 1992;4:121–7 103.

[7] Vaughn DW, Green S, Kalayanaroojm S, et al. Dengue viremia titer, antibody response pattern, and virus serotype correlate with disease severity. J Infect Dis 2000;181:2–9.

[8] Vaughn DW, Green S, Kalayanarooj S, et al. Dengue in the early febrile phase: viremia and antibody responses. J Infect Dis 1997;176:322–30.

[9] Gubler DJ. Dengue and dengue hemorrhagic fever. Clin Microbiol Rev 1998;11: 480–96.

[10] Katzelnick LC, Gresh L, Halloran ME, et al. Antibody-dependent enhancement of severe dengue disease in humans. Science 2017;358:929–32.

[11] Coffey LL, Mertens E, Brehin AC, et al. Human genetic determinants of dengue virus susceptibility. Microbes Infect 2009;11:143–56.

[12] Loke H, Bethell D, Phuong CX, et al. Susceptibility to dengue hemorrhagic fever in vietnam: evidence of an association with variation in the vitamin d receptor and FC gamma receptor IIa genes. Am J Trop Med Hyg 2002;67:102–6.

[13] Sakuntabhai A, Turbpaiboon C, Casadémont I, et al. A variant in the CD209 promoter is associated with severity of dengue disease. Nat Genet 2005;37:507–13.

[14] Pabalan N, Chaisri S, Tabunhan S, Phumyen A, Jarjanazi H, Steiner TS. Associations of DC-sign (CD209) promoter -336G/A polymorphism (rs4804803) with dengue infection: a systematic review and meta-analysis. Acta Trop 2018;177:186–93.

[15] Vejbaesya S, Luangtrakool P, Luangtrakool K, et al. TNF and LTA gene, allele, and extended HLA haplotype associations with severe dengue virus infection in ethnic Thais. J Infect Dis 2009;199:1442–8.

[16] Sierra B, Triska P, Soares P, et al. OSBPL10, rxra and lipid metabolism confer african-ancestry protection against dengue haemorrhagic fever in admixed Cubans. PLoS Pathog 2017:1–21. doi: 10.1371/journal.ppat.1006220.

[17] Stephens HAF. HLA and other gene associations with dengue disease severity. Curr. Top. Microbiol. Immunol 2010;338:99–114. doi: 10.1007/978-3-642-02215-9.

[18] Sánchez-Leyva M, Sánchez-Zazueta JG, Osuna-Ramos JF, et al. Genetic polymorphisms of tumor necrosis factor alpha and susceptibility to dengue virus infection in a mexican population. Viral Immunol 2017 Oct;30(8):615–21.

[19] Dang TN, Naka I, Sa-Ngasang A, et al. Association of BAK1 single nucleotide polymorphism with a risk for dengue hemorrhagic fever. BMC Med Genet 2016;17(1):43.

[20] Xavier-Carvalho C, Cardoso CC, de Souza Kehdy F, Pacheco AG, Moraes MO. Host genetics and dengue fever. Infect Genet Evol 2017 Dec;56:99–110.

[21] Khor CC, Chau TN, Pang J, et al. Genome-wide association study identifies susceptibility loci for dengue shock syndrome at MICB and PLCE1. Nat Genet 2011;43(11):1139–41.

[22] Whitehorn J, Chau TN, Nguyet NM, et al. Genetic variants of MICB and PLCE1 and associations with non-severe dengue. PLoS ONE 2013;8(3):e59067.

[23] Dang TN, Naka I, Sa-ngasang A, Anantapreecha S, Chanama S. A replication study confirms the association of GWAS-identified SNPs at micb and PLCE1 in thai patients with dengue shock syndrome. BMC Med. Genet 2014;15:1–5. doi: 10.1186/1471-2350-15-58.

[24] Lanciotti RS, Calisher CH, Gubler DJ, Chang GJ, Vorndam AV. Rapid detection and typing of dengue viruses from clinical samples by using reverse transcriptase-polymerase chain reaction. J Clin Microbiol 1992;30(3):545–51.

[25] Balmaseda A, Hammond SN, Tellez Y, et al. High seroprevalence of antibodies against dengue virus in a prospective cohort study of schoolchildren in Managua, Nicaragua. Trop Med Intl Health 2006;11:935–42.

[26] Mattia K, Puffer BA, Williams KL, et al. Dengue reporter virus particles for measuring neutralizing antibodies against each of the four dengue serotypes. PLoS ONE 2011;6:e27252.

[27] de Alwis R, de Silva AM. Measuring antibody neutralization of dengue virus (DENV) using a flow cytometry-based technique. Methods Mol Biol 2014;1138:27–39.

[28] Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet 2011;88(1):76–82.

[29] Svishcheva GR, Axenovich TI, Belonogova NM, van Duijn CM, Aulchenko YS. Rapid variance components-based method for whole-genome association analysis. Nat Genet 2012:1166–70.

[30] Visscher PM, Hill WG, Wray NR. Heritability in the genomics era—concepts and misconceptions. Nat Rev Genet 2008;9(4):255–66.

[31] Huy NT, Van Giang T, Thuy DHD, et al. Factors associated with dengue shock syndrome: a systematic review and meta-analysis. Halstead SB, editor. PLoS Negl Trop Dis 2013;7(9):e2412.

[32] Edwards BJ, Haynes C, Levenstien MA, Finch SJ, Gordon D. Power and sample size calculations in the presence of phenotype errors for case/control genetic association studies. BMC Genet 2005;6:18.

AMERICAN SOCIETY FOR MICROBIOLOGY | mBio

| Epidemiology | Research Article

An integrated nationwide genomics study reveals transmission modes of typhoid fever in China

Ye Feng,[1,2] Hang Pan,[3] Beiwen Zheng,[4] Fang Li,[3] Lin Teng,[3] Zhijie Jiang,[3] Mengyao Feng,[3] Xiao Zhou,[3] Xianqi Peng,[3] Xuebin Xu,[5] Haoqiu Wang,[6] Beibei Wu,[7,8] Yonghong Xiao,[4] Stephen Baker,[9] Guoping Zhao,[10,11,12] Min Yue[3,4,13]

AUTHOR AFFILIATIONS See affiliation list on p. 13.

ABSTRACT Typhoid fever, caused by *Salmonella* Typhi (*S*. Typhi), is a life-threatening disease, usually food-borne and commonly associated with international travel. The disease transmission remains endemic in many low- and middle-income countries, representing further hotspots for seeding new global outbreaks. China has historically been affected by typhoid fever, but the respective roles of local transmission and importation remain unknown. Here, we generated a nationwide map of the typhoid burden in China and investigated the associations between typhoid disease, climate and various socioeconomic parameters. To assess transmission dynamics, we sub-sampled *S*. Typhi isolated within China over five decades and sequenced their genomes. The resulting 705 new genomes, placed in context with 5,190 global isolates from 87 countries on six continents, led to the discovery of several predominant inland Chinese clones belonging to the clades 2.1/2.3/3.2/4.3. These clones were associated with multiple introductions from overseas, followed by local expansion. Notably, 4.3.1 isolates from eastern China were not genetically close to those from northwestern China but to the international isolates, indicating their association with international travel. Additional *in vitro* assays showed that 4.3.1 elaborated better intracellular survival, acid tolerance, and desiccation tolerance than other lineages, partially explaining its success. For the first time, we have probed typhoid transmission in China, finding local transmission and importation, which could guide the policy for typhoid control.

IMPORTANCE Typhoid fever is a life-threatening disease caused by *Salmonella enterica* serovar Typhi, resulting in a significant disease burden across developing countries. Historically, China was very much close to the global epicenter of typhoid, but the role of typhoid transmission within China and among epicenter remains overlooked in previous investigations. By using newly produced genomics on a national scale, we clarify the complex local and global transmission history of such a notorious disease agent in China spanning the most recent five decades, which largely undermines the global public health network.

KEYWORDS *Salmonella* Typhi, typhoidal fever, transmission, exploiting genomics, pandemic clone

Salmonella enterica serovar Typhi (*S*. Typhi) is a human-restricted bacterium that causes typhoid fever. The disease is estimated to cause ~21.7 million illnesses and 216,000 deaths annually (1). Typhoid transmission is known to occur through two modes: "long-cycle" transmission, occurs through long-term exposure to contamination in the broader environment (i.e., untreated water or sewage), whereas "short-cycle" transmission, refers to individuals exposed through close interaction with infected individuals via food handling and is often linked to travel (2). Typically, "long-cycle" transmission occurs in endemic countries with inadequate hygiene and safe food/water handling practices,

Editor Julian Parkhill, University of Cambridge, Cambridge, United Kingdom

Address correspondence to Min Yue, myue@zju.edu.cn.

Ye Feng, Hang Pan, and Beiwen Zheng contributed equally to this article. The order of names was determined by the role and research time of contributions to the project.

The authors declare no conflict of interest.

See the funding table on p. 13.

Received 25 May 2023
Accepted 16 August 2023
Published 6 October 2023

Copyright © 2023 Feng et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International license.

including southern Asia, Africa, and South America (1, 3, 4), where the environment is the primary reservoir. In high-income countries with improved sanitation facilities, however, "short-cycle" transmission is dominant, with most cases associated with recent international travel (5), where humans are the main reservoir.

Genetically, *S*. Typhi was traditionally thought to be monophyletic, with limited nucleotide diversity and recombination capability, as revealed by multi-locus sequence typing technology (6–8). With the high resolution of whole-genome sequencing (WGS), the global *S*. Typhi population can be grouped into 5 primary clades and 24 subclades (9, 10). The clade 4.3.1, also termed H58, represents a common genotype that became widely distributed in southern Asia and the Middle East in the 1990s and later spread worldwide (11). Many of the 4.3.1 isolates are characterized by multi-drug resistance (MDR), including resistance against chloramphenicol, ampicillin, and trimethoprim-sulfamethoxazole. MDR was primarily driven by the acquisition of IncHI1 plasmids; however, the more recent integration of additional antimicrobial-resistance (AMR) genes and chromosomal point mutations (12, 13) has conferred resistance to fluoroquinolones, third-generation cephalosporins, and azithromycin in 4.3.1 (14–16).

China has vast geographical, meteorological, and economic diversity across its regions. Historically, typhoid fever was rampant across China. Over the past four decades, dramatic socioeconomic advancement has resulted in the overall incidence of typhoid in China dropping to low levels (17–19). Despite this reduction, the spatiotemporal heterogeneity of typhoid across China is unclear. Whether continuous population migration within China has generated heterogeneity in *S*. Typhi or disseminated endemic clones remains obscure. Additionally, transmission routes within China and potential importations have yet to be defined. To date, few large-scale studies have been conducted except for outbreak reports (20–22), and only 26 *S*. Typhi genome sequences from China are publicly available. Here, we amassed a nationwide collection of *S*. Typhi isolates spanning the last five decades and performed genome sequencing. In combination with global contextual genomes, we performed phylodynamic analysis and investigated the population structure and the scale of internal transmission and importation of *S*. Typhi in China.

RESULTS

Variable typhoid across China

The incidence of typhoid fever has continued to decline since 2004 and has remained constant at 0.8 cases per 100,000 persons per year since 2008 (Fig. 1A). The incidence varied between provinces and was generally found to be higher in northwestern and southern China (Fig. 1B). The incidence was correlated with various climatic factors such as temperature, precipitation, and latitude. However, incidence was not strongly associated with per capita gross domestic product (GDP), the number of outbound tourists, or population density (Fig. 1C; Table S1).

We collected the genomes of 731 *S*. Typhi isolates from 19 provinces across China. A global collection of 5,164 genomes retrieved from the public database was also included to provide a global context. The core genome of these isolates had a size of 3.97 Mb, within which a total of 42,702 core-genome single-nucleotide polymorphisms (cgSNPs) were identified. A Bayesian analysis of population structure typing and GenoTyphi, a 58 SNP-based genotyping scheme (9), were performed, respectively, and produced consistent typing results with each other (Fig. S1).

The genotypic structure of the Chinese isolates was distinct between regions. Northern China was dominated by the primary clade 2, while other regions comprised two or three primary clades (Fig. 2A; Fig. S2). At the subclade level, 2.1 (20.8%), 2.3 (14.5%), 3.2 (21.2%), and 4.3 (9.0%) were dominant in China (Fig. 2B). No obvious clonal replacement was observed in any region during the last two decades (Fig. S3).

The demographic features of patients infected by 4.3 differed from those of other clades, with young people and females accounting for a higher proportion than the elders and males (Fig. S4). The division of the 4.3 patients according to the region further

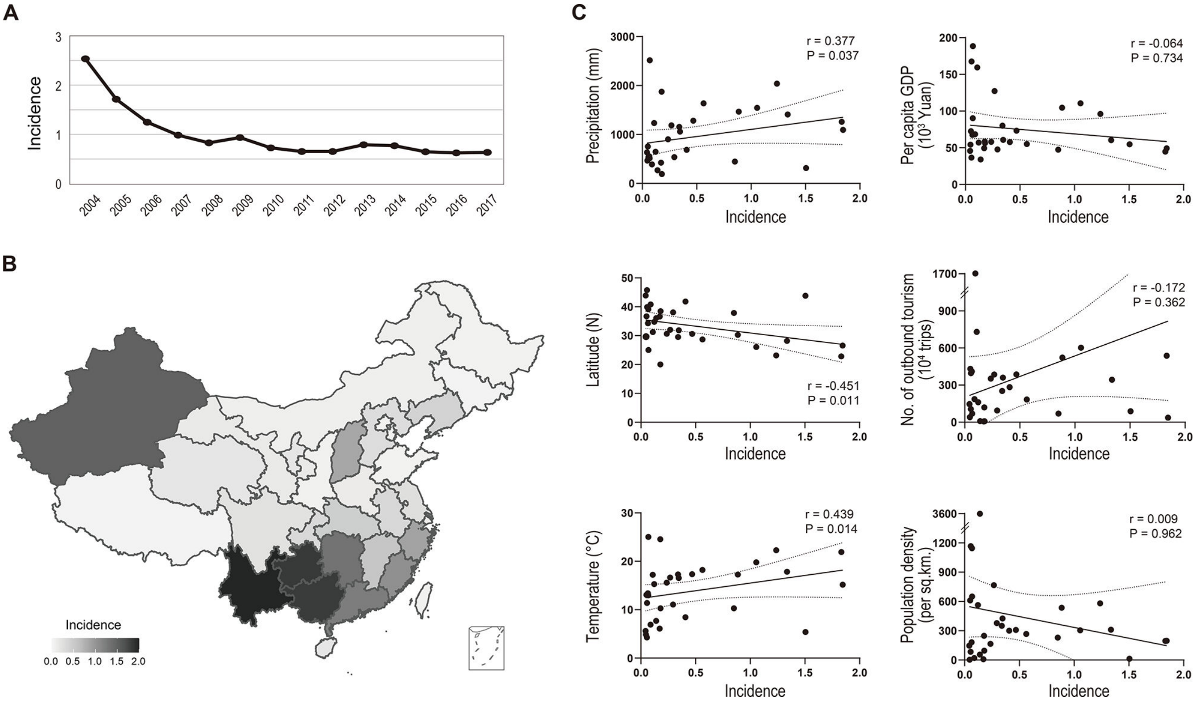


FIG 1 Incidence of typhoid fever in China. (A) Annual incidence (number of cases per 100,000 persons per year) of typhoid fever in China during 2004–2017. (B) The average incidence of typhoid fever in each province of China during 2008–2017. (C) Spearman correlation of climate, social, and economic factors (*y*-axis) with the incidence of typhoid fever (*x*-axis). Each dot indicates a province of China. The data for the plots are appended in Table S1.

showed that the low age of the 4.3 patients was attributable to the patients from eastern China, and the sex bias was attributed to the patients from northwestern China.

International data were obtained from public databases. Overall, Africa, Asia, and Europe were dominated mainly by 4.3 (Fig. 2A and B), although the genotypic composition varied remarkably between and within continents (23).

A low prevalence of AMR *S*. Typhi in China

We further analyzed the distribution of AMR-associated mobile genes and chromosomal mutations, particularly those responsible for resistance to front-line therapy. Only two Chinese *S*. Typhi isolates carried the plasmid-borne *qnr* genes for fluoroquinolone resistance. The *gyr* mutations were present in 12% of isolates and were mainly enriched in the primary clade 4, in which the occurrence was 24%. Three Chinese isolates, which belonged to different clades, carried distinct *bla*$_{\text{CTX-M}}$ gene variants that offered resistance to third-generation cephalosporin. The recently emergent *acrB* mutations that mediate azithromycin resistance, as well as the *mcr*, *tetX*, and carbapenemase-encoding genes, which confer resistance to polymyxin, tigecycline and carbapenem, respectively, were not detected. The rate of plasmid carriage in Chinese isolates was <12% for all primary clades (Fig. 2C), with IncC and IncFIB (pHCM2) as the predominant replicon types (Fig. S5). By comparison, Asian and European isolates were much more resistant than Chinese isolates (Fig. 2C).

Historic “long-cycle” transmission across provinces in China

The clades 2.1, 2.3, 3.2, and 4.3 outnumbered the others. They were frequently recovered from localized provinces over 4 years, indicating their long-term establishment as local reservoirs. To investigate the dates of emergence and geographical transfer of these clades, we verified their temporal structure (Fig. S6) and inferred their timed phylogenies for four major clades (Fig. 3).

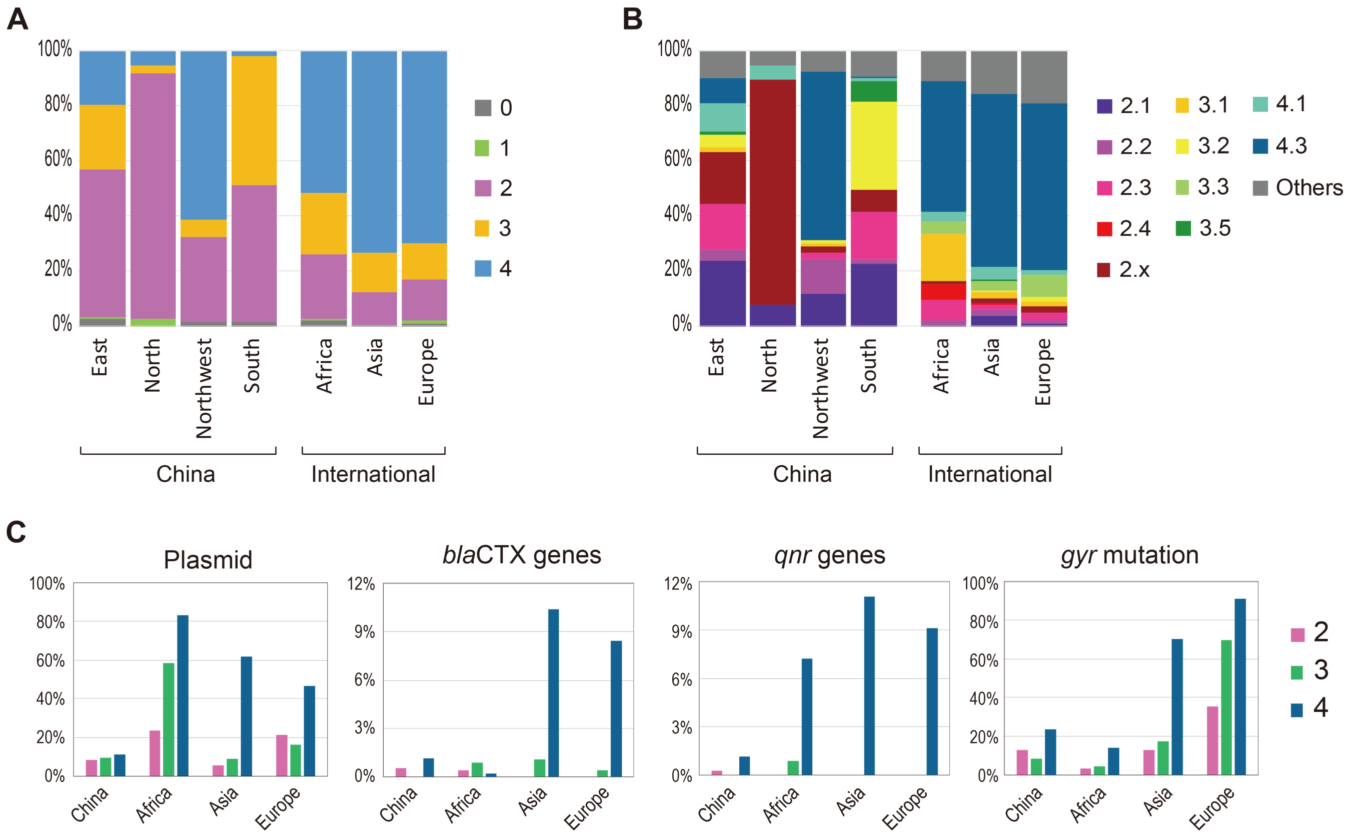


FIG 2 Distinct genotypic and antimicrobial patterns of *S.* Typhi between regions. The *S.* Typhi isolates are grouped into primary clades (A) and subclades (B) by the Genotyphi typing scheme. According to the geographic source, the comparison is made between different parts of China, i.e., eastern (*n* = 144), northern (*n* = 38), northwestern (*n* = 83), southern (*n* = 459), and between various regions over the world, i.e., China (*n* = 731), Africa (*n* = 1,074), Asia (excluding China, *n* = 1,443), and Europe (*n* = 1,942). (C) Comparison of carriage rate of antimicrobial-related genetic determinants.

The profile of geographical origin divided the four clades into two categories (single origin or multiple origin) (Fig. 3A; Fig. S7). The Chinese 3.2 and 4.3 isolates both formed a single branch that separated from the international isolates, indicating a single introduction into China. In contrast, the Chinese 2.1 and 2.3 isolates formed several branches that mixed with international isolates, suggesting that the two clades were introduced into different provinces of China in multiple waves. Notably, 2.3 was found to be repeatedly transmitted into the Guangxi province (in southern China) (Fig. 3A and B). This indicated that, Guangxi, a frontier of Chinese civilization throughout its history and a tourist destination in recent decades, served as a transmission hotspot: the internationally common clones were initially imported into Guangxi and later spread to eastern and northern China (Fig. 3B). Coincidentally, 2.1 and 2.3 entered China much earlier than 3.2 and 4.3. In particular, the establishment of 4.3 as an endemic clone in northwestern China could be traced back to 2002, coinciding with the time that local Muslims frequented the Hajj pilgrimage in the Middle East.

Bayesian estimation of the bacterial population showed a remarkable decline of 2.1, 2.3 and 3.2 since the 1990s (Fig. 3C). Only 4.3 underwent a growth of population, which corresponded to the global dissemination of this lineage.

Recent "short-cycle" transmission within Chinese provinces

As travel-associated cases of *S.* Typhi infections have attracted increasing attention in developed countries (24), we examined putative travel-associated events in China based purely on genome similarity analysis. Except for 4.3, practically all *S.* Typhi isolates in China had their genetically closest relatives collected from the same province. This

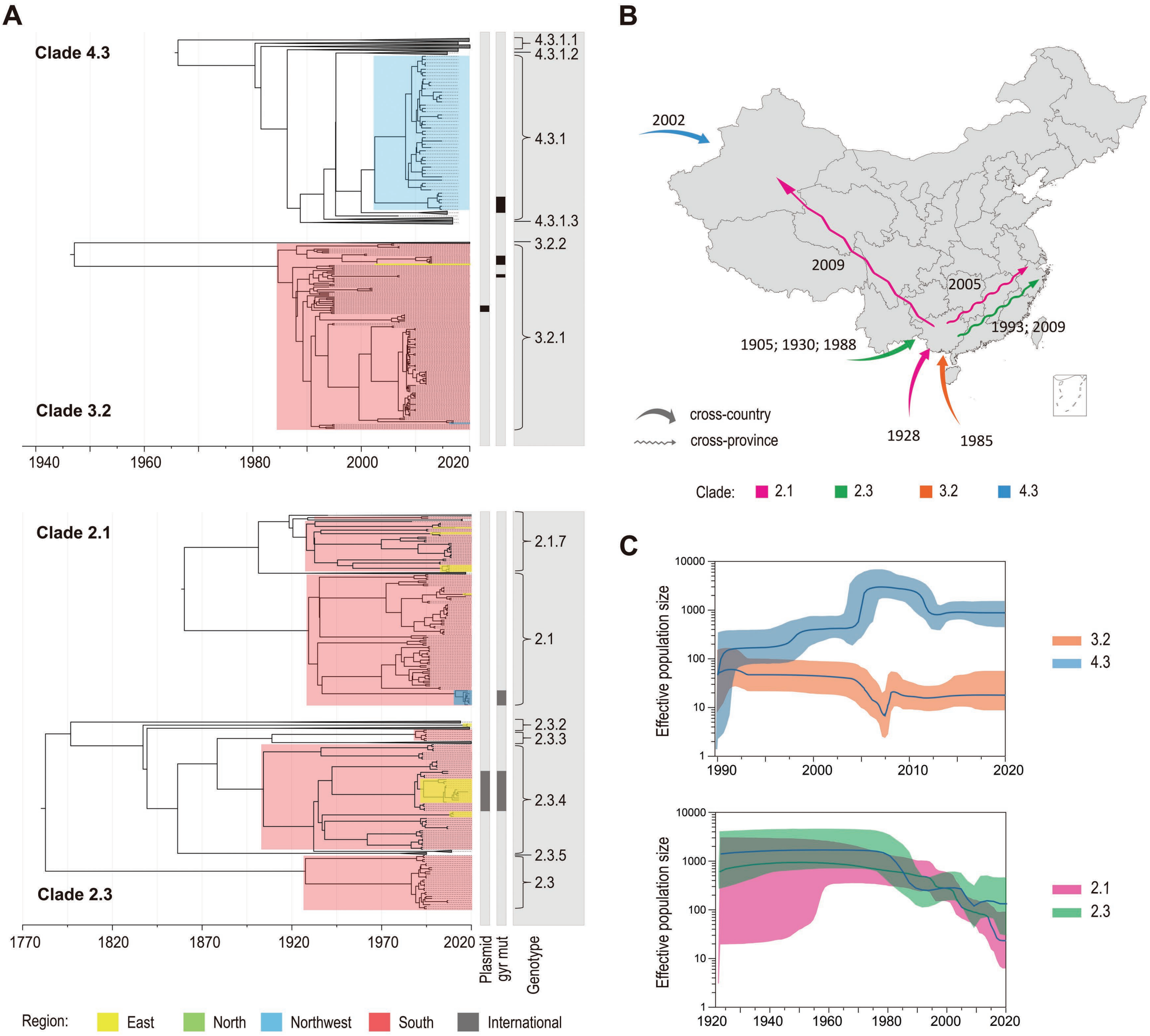


FIG 3 "Long-cycle" transmission of four dominant clades in China. (A) Maximum clade credibility tree (reconstructed using BEAST2) with the antimicrobial-associated metadata. The Chinese *S*. Typhi isolates from four regions are highlighted by boxes in different colors. The branches comprising international isolates are collapsed into triangles in gray. The right columns indicate the presence of plasmid and *gyr* mutations as well as the detailed genotypes. (B) Transmission routes and time inferred by BEAST analysis. Two kinds of arrows show transmission events from abroad and between provinces. The numbers near the arrows indicate the estimated year of the putative transmission events. (C) Bayesian skyline plots show the historical changes in the effective population size of the four clades. Lines and shadings indicate the medians and the 95% HPD intervals of estimated effective population sizes. The *x*-axis shows the time, and the *y*-axis indicates the population size.

finding supports the prevalence of regional endemics in China. Notably, all Chinese 4.3 isolates are 4.3.1. We compared the histogram of the pairwise genetic distance of the isolates within the same province and between different provinces. The former showed a trough between two peaks at five cgSNPs, whereas the latter rose starting from five cgSNPs (Fig. 4A). Based on these findings, we used five cgSNPs as the threshold to determine if the compared isolates belonged to the same episode, resulting in the identification of six instances of cross-province transmission (Fig. 4B). Among them, four involved Guangxi, again demonstrating that Guangxi province has the most frequent within-China transmission events.

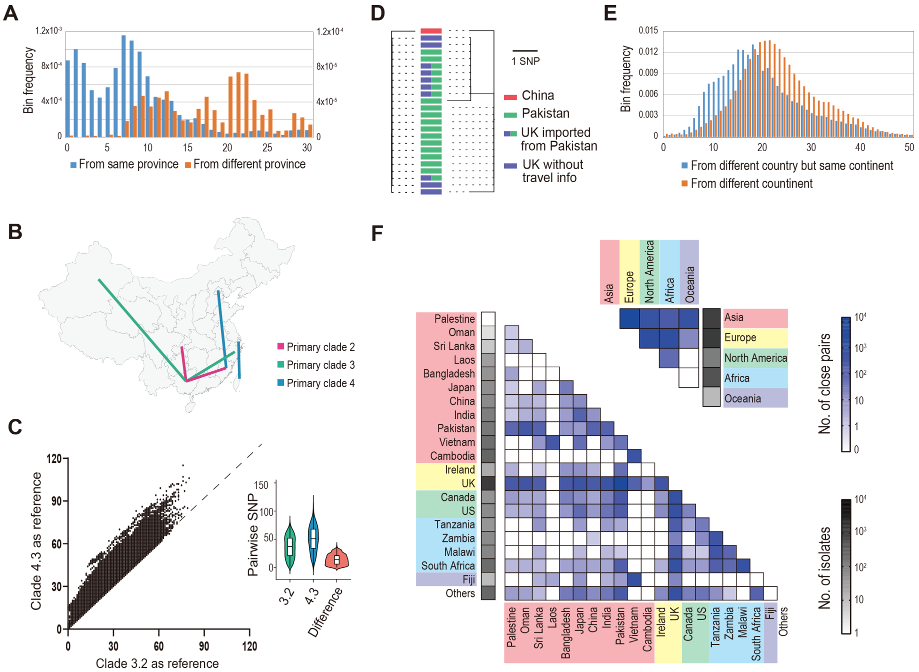


FIG 4 "Short-cycle" (or travel-associated) transmission of *S*. Typhi. (A) Histogram of cgSNP distance for the Chinese *S*. Typhi isolates collected from the same province (left *y*-axis) and different provinces (right *y*-axis). The *x*-axis indicates the number of cgSNPs. (B) Putative transmission events across provinces in China. Each line indicates a pair of *S*. Typhi isolates with pairwise distance <5 cgSNPs, indicating a putative transmission event. (C) Comparison of cgSNP distance between 4.3 isolates derived from different reference genomes as illustrated by a scatter plot and a violin plot, respectively. In the scatter plot, each dot indicates a pair of isolates, with its *x*-axis indicating the cgSNP distance calculated using the genome of strain CT18 (belonging to 3.2) as the reference and its *y*-axis using strain ERL052042 (accession number GCF_001088345, belonging to 4.3) as the reference. In the violin plot, the blue and green violins illustrate the distribution of the two sets of distance; and the red indicates their difference. (D) Example from the Chinese 4.3 isolate 458 (collected from Shanghai city) as well as the isolates of its closest genome match. Using CT18 strain as the reference, the left ML tree shows isolate 458 was one cgSNP distant from the international isolates (from the UK and Pakistan). Using ERL052042 as the reference, the right ML tree shows isolate 458 was one to three cgSNPs distant from these international isolates, but most of the international isolates still have the identical cgSNP profile, i.e., they cannot be distinguished from each other by their pairwise cgSNPs. (E) Histogram of cgSNP distance for 4.3 isolates collected from different countries and continents. The *x*-axis indicates the number of cgSNPs. (F) Genomic relatedness of 4.3 isolates worldwide as revealed by their pairwise cgSNP distance. The heatmaps show the numbers of close cgSNP matches (i.e., <5 cgSNPs) between countries and between continents. Countries with the same shade color belong to the same continents.

The 4.3 isolates collected from eastern China were not genetically closer to those from northwestern China but rather to international isolates (Fig. S7), providing strong evidence of their travel-associated nature. Nevertheless, as the Chinese 4.3 isolates showed an equally small distance from the international isolates from different countries, determining the specific country of origin was challenging. For example, an isolate from Shanghai was diverged from many of the UK and Pakistan isolates by a single cgSNP (Fig. 4D). We attributed this tiny genetic distance to a substantially lower within-clade nucleotide diversity of 4.3 than the other clades (Fig. S8). The calculation of cgSNPs employed the strain CT18 (belonging to 3.2) as the reference, which led to a loss of informative sites within the genomic segments specific to 4.3. Then, we switched to a 4.3 genome as the reference to increase genomic tracing resolution. This alternation expanded the core genome (i.e., the cgSNP searching range) from 3.97 to 4.51 Mb, and the average number of pairwise cgSNPs among 4.3 isolates increased from 36 to 51, indicating a 41.7% increase in resolution (Fig. 4C). Despite the improvement, several of the international isolates in the preceding example still exhibited the identical cgSNP profile (Fig. 4D). Out of the nine UK isolates that resembled the Shanghai isolate, six were imported from Pakistan via international travel (25, 26), explaining their genomic similarity. However, the remaining three lacked information on whether they were imported and from which countries they originated, therefore, their origin in Pakistan cannot be fully confirmed.

Expanded examination to global 4.3 showed the pattern of <5 cgSNP distance for the isolates from different countries of the same continent was, in general, similar to that from different continents, suggesting intercontinental transmission was almost as common as international transmission (Fig. 4E). The most frequent international transmission appears to occur between the UK and Pakistan; the most frequent intercontinental transmission between Europe and Asia (Fig. 4F).

Phenotypic advantages of 4.3.1 over other lineages

To understand the advantage of 4.3.1 as a pandemic clone, we selected strains to represent *S*. Typhi clades of different genetic backgrounds from our Chinese collections and conducted a series of *in vitro* assays (Table S4). In general, the 4.3.1 isolates exhibited significantly better intercellular survival abilities than isolates from the primary clades 0, 2, and 3, within both the human myeloid leukemia cell line (U937) (Fig. 5A) and human monocytic leukemia cell line (THP-1) (Fig. 5B) at 12-h post-infection. Compared with a

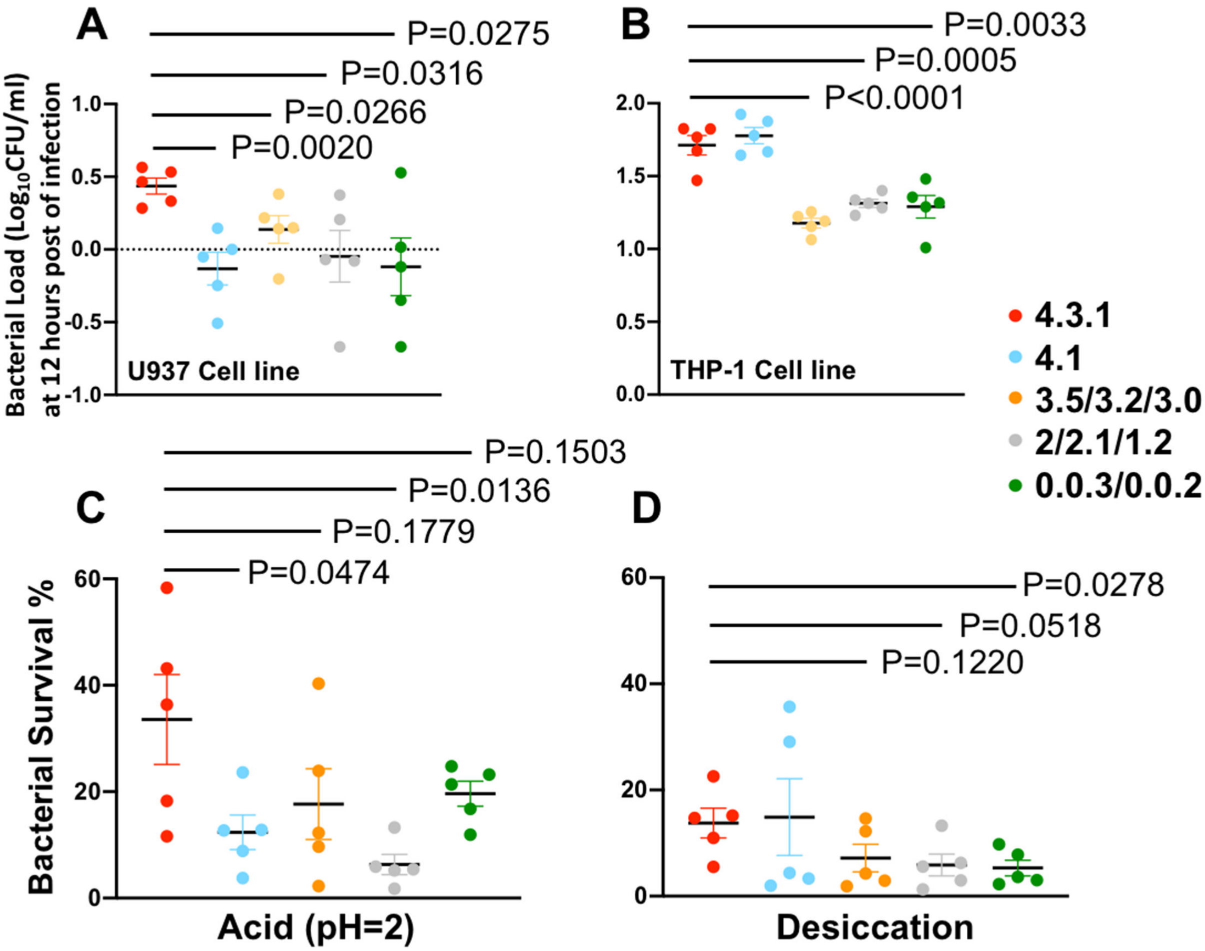


FIG 5 *In vitro* phenotypic comparison between 4.3.1 and other clades. Cellular invasion assays in the U937 cell line (A) and THP-1 cell line (B). The *y*-axis scale represents the $\log_{10}$ value of the ratio of detected bacteria at 12 h when compared with the initial bacteria that successfully invaded. *In vitro* tolerance assays with acid (pH = 2) (C) and desiccation (D). The *y*-axis scale represents the percentage of detected bacteria compared to the initial inoculation. Four groups of clinical isolates (Table S4) were chosen based on their lineage origin and availability in the current collection. The unpaired two-tailed *t* test was performed to assess the difference between groups with the calculated *P* value.

close relative 4.1, 4.3.1 also shows a better survival rate within U937 cells. Regarding tolerance to acid (Fig. 5C) and desiccation (Fig. 5D), we also observed that 4.3.1 isolates have better survival rates, even though the *P* value did not always reach the significant threshold. However, there are no dramatic differences in biofilm formation among lineages under the four examined conditions (Fig. S9). Interestingly, all the isolates show strong biofilm formation capabilities under swine bile salt medium with anaerobic conditions.

DISCUSSION

Here, we conducted, for the first time, a genomic epidemiology study of *S*. Typhi infection in China and found that both the “long-cycle” and “short-cycle” transmission modes commonly co-exist, providing striking regional heterogeneity. Most infections in eastern China resulted from “short-cycle” transmission, as revealed by their close match with international isolate genomes. This finding can be explained by the close connection of the populace in this region with other countries, such as southern Asia. By contrast, isolates from the other provinces displayed explicit geographical restrictions, suggesting that infections are attributable to the long-term circulation of the locally endemic clones. Dated phylogeny analysis further uncovered that endemic Chinese clones were not entirely indigenous but were, and continue to be, imported for several hundreds of years. For example, *S*. Typhi in northwestern China may well have been imported by frequent pilgrims from the Middle East, whereas the *S*. Typhi found in southern China almost certainly originated in adjacent Southeast Asia, a known reservoir for global typhoid fever (4). An interesting question is why *S*. Typhi imported to southern and northwestern China eventually developed into endemic clones, whereas those imported to eastern China only caused sporadic cases. Likewise, 4.3 imported by travelers to European countries has not created endemic disease, whereas importation to Africa via the same route, i.e., traveling public, has established a local dominant clone (27).

It has been recognized that climate and hydrological systems both play pivotal roles in the transmission of typhoid fever (2, 28–30). Northwestern and southern China are warm and damp, suitable for *Salmonella* growth. Furthermore, in the rural areas of these regions, local people utilize river or well water for cooking and cleaning (31, 32), similar to the endemic typhoid regions of South Asia, Southeast Asia, and Africa. As well as the popularity of foreign travel in eastern China, further multi-faceted contributions led to dual transmission modes in China. These may explain the relatively low correlation of typhoidal incidence with the GDP and tourist numbers (Fig. 1C).

In the era of the global village, genome-based epidemiology has been increasingly employed to reconstruct transmission links for many infectious diseases, which is built on the assumption that the natural origin of the disease tends to come from the region with the closest genome match (33, 34). This assumption has a premise that bacterial genomes must show robust geographical clustering, with considerable distance between isolates from different regions. This premise holds for most bacterial pathogens, including *S*. Typhi 2.1, 2.3, and 3.2 in China. By comparison, predicting the geographic origin of 4.3 genomes is much more challenging. The within-clade cgSNP distance is small, and frequent international transfer has distorted geographical clustering. Last but not least, the biased data sources of public databases have impeded global tracking investigations. Many of the developing countries, which are the real sources of typhoid disease, have submitted very few genomic data sets, probably due to the cost of establishing sequencing-based surveillance systems. Whereas the UK holds the largest data set of sequenced isolates, making it more likely to be the country with the closest genome match to any query isolates. Consequently, when the travel history of the UK patients is missing, the UK is easy to be misidentified as the country of origin.

As proposed here, three measures may help address the above issues. First, a hierarchical typing scheme, e.g., using a specific reference genome for 4.3.1 isolates, has been shown to increase the resolution. Second, tracing the exact source country should depend not only on the count of cgSNP distance but also on the topology of the dated phylogeny or even more complicated spatial-temporal modeling that comprises all the closely related isolates (24). Third, well-integrated databases are urgently required, which include not only the site of isolation/sequencing but also the site of infection (or rather the travel history). Last, increasing the density of isolate collection and genome sequencing from developing countries, including China, would effectively avoid sampling bias and help depict an accurate transmission route.

Over the last three decades, intercontinental transmission and circulation have been dominated by 4.3.1 or called H58. In line with this study, we found that the short-cycle international transmission in China was exclusively attributed to 4.3.1. MDR was once thought to drive the worldwide expansion of 4.3.1 (12, 35), but MDR is also present in other clades. Our bile resistance assay refuted the hypothesis that 4.3.1 has higher bile resistance benefiting persistence in the gallbladder in asymptomatic patients, which is in line with a previous report that *S*. Typhi carriage is not restricted to any particular genotype (36). Interestingly, we found an improved ability of 4.3.1 to resist a low pH environment, which might allow easier access to the intestine. We also found that 4.3.1 can better survive and replicate in macrophages. While this ability assists 4.3.1 to elude the host immune response, it might also benefit 4.3.1 living within amoebae in the environment. Similarly, the improved capability of desiccation tolerance may likely contribute to the environmental persistence of 4.3.1. Notably, 4.3.1 may acquire these capabilities in a stepwise manner. Our results showed that 4.3.1 has comparable desiccation tolerance and intracellular survival abilities in THP-1 compared to its close relative 4.1. But 4.3.1 performs better in acid tolerance and survival in U937, suggesting 4.3.1 has evolved the latter two capabilities after it diverged from 4.1. Nevertheless, all the examined results are based on Chinese clinical isolates, further investigations with an extensive diversified collection in the global community are highly warranted.

In conclusion, we provided the largest genomic data set from China for global surveillance and AMR assessment for one of the oldest killer diseases in the world. More importantly, our nationwide study revealed a novel dual-transmission mode with distinct regional features, which also correlated with the climate, and social-economic parameters of those regions in China. A notable historic "long-cycle," mainly endemic in inter-mainland China, is associated with the provinces that require improvements to hygiene, particularly water sanitation. Recent "short-cycle" transmission, mainly in more advanced socioeconomic areas in eastern China, is related to intensified global travel and linked with the MDR pandemic 4.3.1. Furthermore, we provided new biological evidence of tolerance capabilities that may favor typhoid transmission. The gained knowledge in this study would tailor the policy with precise interventions according to distinct transmission hotspots and modes in different regions.

MATERIALS AND METHODS

Data on *S*. Typhi incidence and correlation analysis

The annual incidence of *S*. Typhi infection was obtained from the Data Center of China Public Health Science (https://www.phsciencedata.cn/). The temperature data were retrieved from the website of the National Oceanic and Atmospheric Administration (www.noaa.gov). The precipitation data were recovered from the software WheatA (http://www.wheata.cn/). The latitude data were obtained from the Baidu Baike database (https://baike.baidu.com/), using the provincial capital city to represent the province. The GDP data were retrieved from the OuWei database (https://www.ovo.com.cn/). The tourism data were obtained from the Baidu Wenku database (https://baike.baidu.com/). All the collected data sets are in Table S1. The Spearman correlation between the *S*. Typhi incidence and the variables was calculated with R software v4.0.3.

Bacterial isolates and WGS

A total of 705 *S*. Typhi isolates were recovered between 1960 and 2020, originating from 17 provinces in China (Table S2). The isolates yielded from humans (*n* = 690) or water (*n* = 10). Human samples included blood (*n* = 616), feces (*n* = 36), marrow (*n* = 4), anal swab (*n* = 1), bile (*n* = 1), body fluid (*n* = 1), urine (*n* = 1), and unknown type (*n* = 30). Water samples included effluent (*n* = 6), pipe water (*n* = 3), and well water (*n* = 1). The sources for the remaining samples included environment (toilet, *n* = 2) and unknown source (*n* = 3).

Most of the isolates were collected by several major CDCs around the country, including Zhejiang CDC, Shanghai CDC, and Shandong CDC. The First Affiliated Hospital of Zhejiang University School of Medicine contributed to assembling the Bacterial Resistant Investigation Collaborative System collection, which collected bloodstream patients from 23 tertiary hospitals and 29 non-tertiary hospitals covering 18 provinces of mainland China. These organizations collected the isolates from subordinate local CDCs under an active routine surveillance program. Human samples were collected from patients with clinical diagnosis (*n* = 653), confirmed carrier cases (*n* = 1), and people with unknown status of infection (*n* = 36).

Total genomic DNA was extracted from the *S*. Typhi using the TIANamp genomic DNA extraction Kit (RT405, TIANGEN BIOTECH, China) and subjected to WGS using the Illumina HiSeq 2000 or HiSeq 2500 platforms (Illumina, San Diego, CA, USA) to generate 150 bp paired-end reads. *De novo* assembly of genome sequences was performed using SPAdes v3.12.0 (37). Serotyping was further confirmed by SeqSero2 v1.1.1 and SISTR v1.0.2 (38, 39).

Genotyping and phylogenetic analyses

To provide global context, a total of 5,190 publicly available *S*. Typhi genomes retrieved from the Enterobase *Salmonella* database and NCBI genome database (until 9 May 2021, including 26 genomes of Chinese isolates and 5,164 international isolates with the genomes) were included for analysis in addition to the genomes sequenced in this study (Tables S2 and S3). The cgSNPs were classified by mapping the assemblies of all query genomes to the reference sequence of *S*. Typhi strain CT18 (accession no: AL513382) with SNIPPY software v4.4.4 (40). The recombinant regions were removed using Gubbins v2.4.1, inset in SNIPPY (41). From the identified SNPs, a subset of 68 was assigned *S*. Typhi isolates to previously defined lineages according to the existing extended *S*. Typhi genotyping framework with GenoTyphi v1.9.1 (9). The genomes belonging to 4.3 were also mapped to the reference sequence of 4.3 strain ERL052042 (accession no: GCF_001088345) to obtain a second set of cgSNP using the same analytical procedures as the above and below methods.

Concatenated cgSNPs were aligned to build maximum likelihood (ML) phylogenetic trees using IQ-Tree v2.2.0 with default parameters (42). The ML trees were displayed and annotated using iTOL v6 (43). The population structure of the global *S*. Typhi collection was also analyzed by hierarchical Bayesian Analysis of Population Structure v6.0 (44), a phylogeny-free population genetics approach; six levels of clustering were performed in the hierarchy, and a prior upper boundary of 80 clusters was set.

Pairwise genetic distances (the difference in the number of cgSNPs) within and between SCs were calculated from the cgSNP alignment using the dist.dna function in the R package ape v5.4.1.

Temporal and phylogeographical analysis

To verify the temporal structure of the dominant four clades in China, we first used TempEst v1.5.1 to assess temporal structure by conducting a regression analysis of the root-to-tip branch distances of the ML tree as a function of sampling date (45). A date randomization test was also conducted by the TipDatingBeast v1.1 R package (46), in which 20 times of analysis were repeated (47). Then, the four clades were subjected to model-based most recent common ancestor dating and phylogeographical reconstruction using Bayesian evolutionary analysis by sampling trees (BEAST) v2.6.7 (48). To understand international transmission events, the representatives of international isolates were also included in this analysis. We identified clusters of closely related isolates with the same sample source, isolation year, and country of isolation and differed by less than five cgSNPs. One representative isolate was randomly selected from each cluster and was kept in the analysis.

The BEAST analysis was performed on alignments of concatenated cgSNPs following previously described procedures (11). Briefly, the GTR+Γ4 substitution model, Bayesian skyline model, and relaxed log-normal molecular clock were chosen; the chain length was 100 million, with sampling at every 10,000 iterations. Maximum clade credibility trees were generated using TreeAnnotator v2.6.7. Figtree v1.4.3 was used to annotate and visualize the tree. The effective population size was estimated by using Tracer v1.7.1.

Identification of AMR-associated determinants

ABRicate v0.8 and the CGE ResFinder database were used to investigate AMR-associated genes (49). Plasmid replicons were identified using ABRicate v0.8 and PlasmidFinder database (50). Point mutations in the quinolone-resistance determining region of genes *gyrA*, *gyrB*, and *parC*; and the *acrB*- mutations associated with azithromycin resistance were detected using ResFinder v4.1.0 and PointFinder database (49).

Rational choice of Chinese bacterial strains for *in vivo* assays

The strains used in this study are listed in Table S4. These strains were chosen to better reflect significant Chinese clades, with the purpose of comparative analysis for pandemic clone 4.3.1. Among three major primary clades (2, 3, and 4), two to three additional major genotypes of each primary clade were considered, including 2.1, 2, and 2.3.4 from primary clade 2, 3.2.1, 3.5, and 3.0.1 from primary clade 3, and 4.3.1 and 4.1 from primary clade 4. For other minor primary clades, two genotypes with clearly known classification names (0.0.3 and 0.0.2) were selected for primary clade 0, and we used the only available strain representing primary clade 1. For lineage 4.3.1, each of its three genotypes (4.3.1, 4.3.1.2, and 4.3.1.1P1) was highly linked to a geographical region of China: subtype 4.3.1 was mainly from Northwestern China (51/53, 96.2%), lineage 4.3.1.1 was mainly from Eastern China (6/6, 100%), and lineage 4.3.1.2 was mainly from Eastern China (5/5, 100%). Additionally, diversities of years of bacterial isolation and ages/sexes of corresponding patients were both fully considered for a better-represented isolate. Notably, our choice of strains is highly limited to our Chinese collection. The above strains were recovered and routinely grown in Luria Broth (LB) medium.

Acid stress assay

Overnight cultures of 26 strains were transferred to fresh LB medium at a ratio of 1:100 and cultured at 37°C, 220 r/min to OD_{600nm} 0.4–0.5 (10^9 CFU/mL). Plate-count (PC) method was used to record the initial bacterial concentration. Appropriate diluent (10 µL) of the original inoculum was streaked on an LB plate, and incubated at 37°C for 14–18 h (three replicates were performed for each sample). NaCl (Sinopharm Chemical Reagent Co., Ltd), NaOH, and HCl (AR, SCR/HuShi) were used to prepare physiological saline (pH = 2). A 10 µL resuspension solution mentioned above was 10× gradient diluted in physiological saline (pH = 2) to 10^7 CFU/mL and then placed at 37°C for 10 min. The plate counting method was used for recording the number of viable bacteria. The survival rate of the bacteria was calculated (the number of viable numbers/the number of initial bacteria × 100%). Data were analyzed using GraphPad Prism v9.0 software. The unpaired two-tailed *t* test was performed to assess the difference between groups.

Desiccation suppression assay

Overnight culture (16 h at 37°C 220 r/min) of 26 strains was prepared and adjusted to OD_{600nm} value 1.0 (10^9 CFU/mL). A total of 50 µL inoculum was added to a 96-well plate. And the plate was placed in a sealed box containing saturated potassium acetate (Sinopharm Chemical Reagent Co., Ltd) at 37°C for 24 h. After that, the plate counting method was performed (three replicates for each strain) to calculate the survival rate of the bacteria. The unpaired two-tailed *t* test was performed to assess the difference between groups.

Biofilm assay

Biofilm assay was performed in a similar way to our previously reported (51, 52). As mentioned in the desiccation suppression assay, overnight culturing was carried out in LB broth. Bacteria were transferred at a ratio of 1:100 into LB medium supplemented with Swine Bile Salt (0.5%), Ox Bile Salt (0.1%), and No.3 Bile Salt (0.1%), respectively. These bile salts were purchased in Shanghai Bo Microbiology Technology Co., Ltd. After thoroughly mixing, 200 µL inoculum was seeded in each 96-well plate in quintuplicate. *Escherichia coli* strain ATCC25922 and *P. aeruginosa* ATCC27853 were used as the positive control. The inoculated plate was incubated at 28°C for 24 h under aerobic and anaerobic conditions, respectively. After 24 h, the wells of the inoculated plate were washed gently with ddH_2O for three times. After drying, each well was stained with 200 µL of crystal violet (Sinopharm Chemical Reagent Co., Ltd) at room temperature for 20 min. After staining, each well was washed gently with ddH_2O for three times and 200 µL f 75% ethanol (Sinopharm Chemical Reagent Co., Ltd) was added. The dye was wholly dissolved and decolorized by shaking for 20 min. Before measuring results, the product of decolorization was transferred to a new microtiter plate. OD_{600nm} value of each stained well was measured using a microtiter-plate reader of the Tecan spectrophotometer. The unpaired two-tailed *t* test was performed to assess the difference between groups.

Intracellular survival assay

Cellular assays referred to methods previously reported (53, 54). RPMI1640 (Gibco) containing 5% fetal bovine serum (FBS; purchased from Biowest) was applied as a cell culture medium for recovering and culturing. Cells were grown in an incubator maintained at 37°C with 5% CO_2 and passaged to the fifth generation. The THP-1 cells were diluted to 5×10^4 cells/mL by RPMI1640 supplemented with 5% FBS and 50 ng/mL PMA for 48 h. U937 cells were diluted to 5×10^4 cells/mL by RPMI1640 supplemented with 5% FBS and 25 ng/mL PMA for 48 h to prepare the cell lines. The culture medium was discarded and replaced with RPMI1640 (containing 5% FBS) for an additional 12 h culture.

The intracellular survival assay was performed as follows: (i) the overnight culturing was carried out in LB broth as mentioned in the desiccation suppression assay; (ii) the inoculum was diluted with RPMI1640, and a volume of 200 µL was transferred to the well for infecting cells with an MOI of 50:1 for 1 h at 37°C, while plate counting method was performed for recording concentration of initial bacteria in the inoculum; (iii) the cells were washed three times with warm RPMI1640 to remove extracellular bacteria; (4) gentamicin was added to the RPMI1640 medium (containing 5% FBS) at a final concentration of 100 µg/mL and the culture was incubated for 1 h to kill extracellular bacteria; (5) RPMI1640 medium (containing 5% FBS) supplemented with 10 g/mL gentamicin (Sangon Biotech) was used for the new culture after discarding the old medium, and the new culture was incubated at 37°C with 5% CO_2 for 0 and 12 h, respectively, while the cell lysis were subjected to plate couting at each time point; and (6) the survival rate of the bacteria was calculated (the number of viable bacteria at 12 h/the number of initial bacteria that successfully invaded at 0 h × 100%). The unpaired two-tailed t test was performed to assess the difference between groups.

ACKNOWLEDGMENTS

This work was supported by the Zhejiang Provincial Key R&D Program of China (2023C03045), the National Program on Key Research Project of China (2022YFC2604201), the European Union's Horizon 2020 Research and Innovation Program under grant agreement no. 861917 – SAFFI, the National Natural Science Foundation of China (82072241), and the Key Research and Development Program of Hangzhou (202203A08).

AUTHOR AFFILIATIONS

[1]Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou, China
[2]Institute of Translational Medicine, Zhejiang University School of Medicine, Hangzhou, China
[3]Department of Veterinary Medicine, Zhejiang University College of Animal Sciences, Hangzhou, China
[4]State Key Laboratory for Diagnosis and Treatment of Infectious Diseases, National Clinical Research Center for Infectious Diseases, National Medical Center for Infectious Diseases, The First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, China
[5]Shanghai Municipal Center for Disease Control and Prevention, Shanghai, China
[6]Hangzhou Center for Disease Control and Prevention, Hangzhou, China
[7]Zhejiang Province Center for Disease Control and Prevention, Hangzhou, China
[8]School of Public Health and Managemet, Wenzhou Medical University, Wenzhou, Zhejiang, China
[9]University of Cambridge School of Clinical Medicine, Cambridge Biomedical Campus, Cambridge, United Kingdom
[10]School of Life Science, Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Hangzhou, China
[11]CAS Key Laboratory of Synthetic Biology, Institute of Plant Physiology and Ecology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China
[12]Department of Microbiology and Microbial Engineering, School of Life Sciences, Fudan University, Shanghai, China
[13]Hainan Institute of Zhejiang University, Sanya, China

AUTHOR ORCIDs

Ye Feng http://orcid.org/0000-0002-6551-1560
Hang Pan http://orcid.org/0000-0002-6372-2479
Min Yue http://orcid.org/0000-0002-6787-0794

FUNDING

Funder	Grant(s)	Author(s)
National Program on Key Research Project of China	2022YFC2604201	Min Yue
European Union's Horizon 2020 Research and Innovation Programme-SAFFI	861917	Min Yue

AUTHOR CONTRIBUTIONS

Ye Feng, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review and editing | Hang Pan, Formal analysis, Investigation, Methodology, Software, Writing – original draft | Beiwen Zheng, Data curation, Methodology, Resources | Fang Li, Data curation, Investigation, Methodology, Visualization | Lin Teng, Data curation, Investigation, Methodology | Zhijie Jiang, Investigation, Methodology, Validation | Mengyao Feng, Investigation, Validation | Xiao Zhou,

Investigation | Xianqi Peng, Investigation | Xuebin Xu, Data curation | Haoqiu Wang, Data curation | Beibei Wu, Data curation | Yonghong Xiao, Data curation | Stephen Baker, Writing – review and editing | Guoping Zhao, Writing – review and editing | Min Yue, Conceptualization, Formal analysis, Funding acquisition, Project administration, Supervision, Writing – review and editing

DATA AVAILABILITY

Raw sequence data/genome assemblies are available in the NCBI SRA database under accession PRJNA875300, the *Salmonella* public database of Enterobase (55), or the NCBI Genome database. Tables S2 and S3 list the accession number for each isolate used in this study.

ETHICAL STATEMENTS

Protocols were approved by the Research Ethics Committees of the first affiliated hospital, Zhejiang University College of Medicine (2018-752) and Zhejiang Provincial Center for Disease Control and Prevention (2019-014).

ADDITIONAL FILES

The following material is available online.

Supplemental Material

Fig. S2 (mBio01333-23-S0001.tif). Geographic analysis.
Fig. S3 (mBio01333-23-S0002.tif). Clonal replacement.
Fig. S4 (mBio01333-23-S0003.tif). Age and sex analysis.
Fig. S5 (mBio01333-23-S0004.tif). Plasmid analysis.
Fig. S6 (mBio01333-23-S0005.pdf). Assessment of the temporal structure of clades.
Fig. S7 (mBio01333-23-S0006.pdf). ML tree.
Fig. S8 (mBio01333-23-S0007.tif). Pairwise SNP.
Fig. S9 (mBio01333-23-S0008.tif). Biofilm analysis.
Supplemental material (mBio01333-23-S0009.pdf). Supplemental legends and Fig. S1.
Supplemental tables (mBio01333-23-S0010.xlsx). Tables S1 to S4.

REFERENCES

1. Stanaway JD, Reiner RC, Blacker BF, Goldberg EM, Khalil IA, Troeger CE, Andrews JR, Bhutta ZA, Crump JA, Im J, Marks F, Mintz E, Park SE, Zaidi AKM, Abebe Z, Abejie AN, Adedeji IA, Ali BA, Amare AT, Atalay HT, Avokpaho E, Bacha U, Barac A, Bedi N, Berhane A, Browne AJ, Chirinos JL, Chitheer A, Dolecek C, El Sayed Zaki M, Eshrati B, Foreman KJ, Gemechu A, Gupta R, Hailu GB, Henok A, Hibstu DT, Hoang CL, Ilesanmi OS, Iyer VJ, Kahsay A, Kasaeian A, Kassa TD, Khan EA, Khang Y-H, Magdy Abd El Razek H, Melku M, Mengistu DT, Mohammad KA, Mohammed S, Mokdad AH, Nachega JB, Naheed A, Nguyen CT, Nguyen HLT, Nguyen LH, Nguyen NB, Nguyen TH, Nirayo YL, Pangestu T, Patton GC, Qorbani M, Rai RK, Rana SM, Ranabhat CL, Roba KT, Roberts NLS, Rubino S, Safiri S, Sartorius B, Sawhney M, Shiferaw MS, Smith DL, Sykes BL, Tran BX, Tran TT, Ukwaja KN, Vu GT, Vu LG, Weldegebreal F, Yenit MK, Murray CJL, Hay SI. 2019. The global burden of typhoid and paratyphoid fevers: a systematic analysis for the Global Burden of Disease Study 2017. Lancet Infect Dis 19:369–381. https://doi.org/10.1016/S1473-3099(18)30685-6
2. Gauld JS, Olgemoeller F, Heinz E, Nkhata R, Bilima S, Wailan AM, Kennedy N, Mallewa J, Gordon MA, Read JM, Heyderman RS, Thomson NR, Diggle PJ, Feasey NA. 2022. Spatial and genomic data to characterize endemic typhoid transmission. Clin Infect Dis 74:1993–2000. https://doi.org/10.1093/cid/ciab745
3. Meiring JE, Shakya M, Khanam F, Voysey M, Phillips MT, Tonks S, Thindwa D, Darton TC, Dongol S, Karkey A, Zaman K, Baker S, Dolecek C, Dunstan SJ, Dougan G, Holt KE, Heyderman RS, Qadri F, Pitzer VE, Basnyat B, Gordon MA, Clemens J, Pollard AJ, STRATAA Study Group. 2021. Burden of enteric fever at three urban sites in Africa and Asia: a multicentre population-based study. Lancet Glob Health 9:e1688–e1696. https://doi.org/10.1016/S2214-109X(21)00370-3
4. Carey ME, MacWright WR, Im J, Meiring JE, Gibani MM, Park SE, Longley A, Jeon HJ, Hemlock C, Yu AT, Soura A, Aiemjoy K, Owusu-Dabo E, Terferi M, Islam S, Lunguya O, Jacobs J, Gordon M, Dolecek C, Baker S, Pitzer VE, Yousafzai MT, Tonks S, Clemens JD, Date K, Qadri F, Heyderman RS, Saha SK, Basnyat B, Okeke IN, Qamar FN, Voysey M, Luby S, Kang G, Andrews J, Pollard AJ, John J, Garrett D, Marks F. 2020. The surveillance for enteric fever in Asia Project (SEAP), Severe Typhoid fever surveillance in Africa (SETA), surveillance of enteric fever in India (SEFI), and strategic typhoid alliance across Africa and Asia (STRATAA) population-based enteric fever studies: a review of methodological similarities and differences. Clin Infect Dis 71:S102–S110. https://doi.org/10.1093/cid/ciaa367
5. Im J, Islam MT, Ahmmed F, Kim DR, Islam Khan A, Zaman K, Ali M, Marks F, Qadri F, Kim J, Clemens JD. 2021. Can existing improvements of water, sanitation, and hygiene (WASH) in urban slums reduce the burden of typhoid fever in these settings? Clin Infect Dis 72:e720–e726. https://doi.org/10.1093/cid/ciaa1429
6. Holt KE, Parkhill J, Mazzoni CJ, Roumagnac P, Weill F-X, Goodhead I, Rance R, Baker S, Maskell DJ, Wain J, Dolecek C, Achtman M, Dougan G. 2008. High-throughput sequencing provides insights into genome variation and evolution in *Salmonella* Typhi. Nat Genet 40:987–993. https://doi.org/10.1038/ng.195
7. Kidgell C, Reichard U, Wain J, Linz B, Torpdahl M, Dougan G, Achtman M. 2002. *Salmonella* Typhi, the causative agent of typhoid fever, is

approximately 50,000 years old. Infect Genet Evol 2:39–45. https://doi.org/10.1016/s1567-1348(02)00089-8

8. Pan H, Jia C, Paudyal N, Li F, Mao J, Liu X, Dong C, Zhou K, Liao X, Gong J, Fang W, Li X, Kehrenberg C, Yue M. 2022. Comprehensive assessment of subtyping methods for improved surveillance of foodborne *Salmonella*. Microbiol Spectr 10:e0247922. https://doi.org/10.1128/spectrum.02479-22
9. Wong VK, Baker S, Connor TR, Pickard D, Page AJ, Dave J, Murphy N, Holliman R, Sefton A, Millar M, Dyson ZA, Dougan G, Holt KE, International Typhoid Consortium. 2016. An extended genotyping framework for *Salmonella* enterica serovar typhi, the cause of human typhoid. Nat Commun 7:12827. https://doi.org/10.1038/ncomms12827
10. Dyson ZA, Holt KE. 2021. Five years of GenoTyphi: updates to the global *Salmonella* typhi genotyping framework. J Infect Dis 224:S775–S780. https://doi.org/10.1093/infdis/jiab414
11. Wong VK, Baker S, Pickard DJ, Parkhill J, Page AJ, Feasey NA, Kingsley RA, Thomson NR, Keane JA, Weill F-X, Edwards DJ, Hawkey J, Harris SR, Mather AE, Cain AK, Hadfield J, Hart PJ, Thieu NTV, Klemm EJ, Glinos DA, Breiman RF, Watson CH, Kariuki S, Gordon MA, Heyderman RS, Okoro C, Jacobs J, Lunguya O, Edmunds WJ, Msefula C, Chabalgoity JA, Kama M, Jenkins K, Dutta S, Marks F, Campos J, Thompson C, Obaro S, MacLennan CA, Dolecek C, Keddy KH, Smith AM, Parry CM, Karkey A, Mulholland EK, Campbell JI, Dongol S, Basnyat B, Dufour M, Bandaranayake D, Naseri TT, Singh SP, Hatta M, Newton P, Onsare RS, Isaia L, Dance D, Davong V, Thwaites G, Wijedoru L, Crump JA, De Pinna E, Nair S, Nilles EJ, Thanh DP, Turner P, Soeng S, Valcanis M, Powling J, Dimovski K, Hogg G, Farrar J, Holt KE, Dougan G. 2015. Phylogeographical analysis of the dominant multidrug-resistant H58 clade of *Salmonella* typhi identifies inter- and intracontinental transmission events. Nat Genet 47:632–639. https://doi.org/10.1038/ng.3281
12. Holt KE, Phan MD, Baker S, Duy PT, Nga TVT, Nair S, Turner AK, Walsh C, Fanning S, Farrell-Ward S, Dutta S, Kariuki S, Weill F-X, Parkhill J, Dougan G, Wain J. 2011. Emergence of a globally dominant IncHI1 plasmid type associated with multiple drug resistant typhoid. PLoS Negl Trop Dis 5:e1245. https://doi.org/10.1371/journal.pntd.0001245
13. Hendriksen RS, Leekitcharoenphon P, Lukjancenko O, Lukwesa-Musyani C, Tambatamba B, Mwaba J, Kalonda A, Nakazwe R, Kwenda G, Jensen JD, Svendsen CA, Dittmann KK, Kaas RS, Cavaco LM, Aarestrup FM, Hasman H, Mwansa JCL. 2015. Genomic signature of multidrug-resistant *Salmonella enterica* serovar typhi isolates related to a massive outbreak in Zambia between 2010 and 2012. J Clin Microbiol 53:262–272. https://doi.org/10.1128/JCM.02026-14
14. Argimón S, Nagaraj G, Shamanna V, Sravani D, Vasanth AK, Prasanna A, Poojary A, Bari AK, Underwood A, Kekre M, Baker S, Aanensen DM, Lingegowda RK. 2022. Circulation of third-generation cephalosporin resistant *Salmonella* Typhi in Mumbai, India. Clin Infect Dis 74:2234–2237. https://doi.org/10.1093/cid/ciab897
15. Carey ME, Jain R, Yousuf M, Maes M, Dyson ZA, Thu TNH, Nguyen Thi Nguyen T, Ho Ngoc Dan T, Nhu Pham Nguyen Q, Mahindroo J, Thanh Pham D, Sandha KS, Baker S, Taneja N. 2021. Spontaneous emergence of azithromycin resistance in independent lineages of *Salmonella* Typhi in northern India. Clin Infect Dis 72:e120–e127. https://doi.org/10.1093/cid/ciaa1773
16. Klemm EJ, Shakoor S, Page AJ, Qamar FN, Judge K, Saeed DK, Wong VK, Dallman TJ, Nair S, Baker S, Shaheen G, Qureshi S, Yousafzai MT, Saleem MK, Hasan Z, Dougan G, Hasan R. 2018. Emergence of an extensively drug-resistant *Salmonella enterica* serovar typhi clone harboring a promiscuous plasmid encoding resistance to fluoroquinolones and third-generation cephalosporins. mBio 9:e00105-18. https://doi.org/10.1128/mBio.00105-18
17. Song Q, Yang Y, Lin W, Yi B, Xu G. 2017. Epidemiological characteristics and clinical treatment outcome of typhoid fever in Ningbo, China, 2005-2014: pulsed-field GEL electorophoresis results revealing great proportion of common transmission sources. Jpn J Infect Dis 70:513–517. https://doi.org/10.7883/yoken.JJID.2016.434
18. Wang J-F, Wang Y, Zhang J, Christakos G, Sun J-L, Liu X, Lu L, Fu X-Q, Shi Y-Q, Li X-M. 2013. Spatiotemporal transmission and determinants of typhoid and paratyphoid fever in Hongta District, Yunnan Province, China. PLoS Negl Trop Dis 7:e2112. https://doi.org/10.1371/journal.pntd.0002112
19. Yue M, Bai L, Song H, Fang W. 2021. Impacts of microbial food safety in China and beyond. Foodborne Pathog Dis 18:508–509. https://doi.org/10.1089/fpd.2021.29015.int
20. Yan M, Li X, Liao Q, Li F, Zhang J, Kan B. 2016. The emergence and outbreak of multidrug-resistant typhoid fever in China. Emerg Microbes Infect 5:e62. https://doi.org/10.1038/emi.2016.62
21. Hu B, Hou P, Teng L, Miao S, Zhao L, Ji S, Li T, Kehrenberg C, Kang D, Yue M. 2022. Genomic investigation reveals a community typhoid outbreak caused by contaminated drinking water in China, 2016. Front Med (Lausanne) 9:753085. https://doi.org/10.3389/fmed.2022.753085
22. Wang Y, Lu D, Jin Y, Wang H, Lyu B, Zhang X, Huang Y, Shu G, Liu B, Lin C, Zhao H, Zhao M, Shen L, Gao Z, Zhang D, Wang Q, Qu M, Jia L. 2022. Extensively drug-resistant (XDR) *Salmonella* Typhi outbreak by waterborne infection - Beijing municipality, China, January-February 2022. China CDC Wkly 4:254–258. https://doi.org/10.46234/ccdcw2022.062
23. Carey ME, Dyson ZA, Ingle DJ, Amir A, Aworh MK, Chattaway MA, Chew KL, Crump JA, Feasey NA, Howden BP, Keddy KH, Maes M, Parry CM, Puyvelde SV, Webb HE, Afolayan AO, Anandan S, Andrews JR, Ashton PM, Basnyat B, Bavdekar A, Bogoch II, Clemens JD, da Silva KE, De A, de Ligt J, Diaz Guevara PL, Dolecek C, Dutta S, Watkins LF, Garrett DO, Godbole G, Gordon MA, Greenhill AR, Griffin C, Gupta M, Hendricksen R, Heyderman RS, Hooda Y, Hormazabal JC, Ikhimiukor OO, Iqbal J, Jacob JJ, Jenkins C, Jinka DR, John J, Kang G, Kanteh A, Kapil A, Karkey A, Kariuki S, Kingsley RA, Koshy RM, Lauer AC, Levine MM, Lingegowda RK, Luby SP, Mackenzie GA, Mashe TA, Msefula C, Mutreja A, Nagaraj G, Nagaraj S, Nair S, Naseri TK, Nimarota-Brown S, Njamkepo E, Okeke IN, Perumal SPB, Pollard AJ, Pragasam AK, Qadri F, Qamar FN, Rahman SIA, Rambocus SD, Rasko DA, Ray P, Robins-Browne R, Rongsen-Chandola T, Rutanga JP, Saha SK, Saha S, Saigal K, Sajib MSI, Seidman JC, Shakya J, Shamanna V, Shastri J, Shrestha R, Sia S, Sikorski MJ, Singh A, Smith AM, Tagg KA, Tamrakar D, Tanmoy AM, Thomas M, Thomas MS, Thomsen R, Thomson NR, Tupua S, Vaidya K, Valcanis M, Veeraraghavan B, Weill F-X, Wright J, Dougan G, Argimón S, Keane JA, Aanensen DM, Baker S, Holt KE, Global Typhoid Genomics Consortium Group Authorship. 2022. Global diversity and antimicrobial resistance of typhoid fever pathogens: insights from 13,000 *Salmonella* typhi genomes. medRxiv. https://doi.org/10.1101/2022.12.28.22283969
24. da Silva KE, Tanmoy AM, Pragasam AK, Iqbal J, Sajib MSI, Mutreja A, Veeraraghavan B, Tamrakar D, Qamar FN, Dougan G, Bogoch I, Seidman JC, Shakya J, Vaidya K, Carey ME, Shrestha R, Irfan S, Baker S, Luby SP, Cao Y, Dyson ZA, Garrett DO, John J, Kang G, Hooda Y, Saha SK, Saha S, Andrews JR. 2022. The international and intercontinental spread and expansion of antimicrobial-resistant *Salmonella* Typhi: a genomic epidemiology study. Lancet Microbe 3:e567–e577. https://doi.org/10.1016/S2666-5247(22)00093-3
25. Nair S, Chattaway M, Langridge GC, Gentle A, Day M, Ainsworth EV, Mohamed I, Smith R, Jenkins C, Dallman TJ, Godbole G. 2021. ESBL-producing strains isolated from imported cases of enteric fever in England and Wales reveal multiple chromosomal integrations of *bla*$_{\text{CTX-M}}$-15 in XDR *Salmonella* Typhi. J Antimicrob Chemother 76:1459–1466. https://doi.org/10.1093/jac/dkab049
26. Chattaway MA, Gentle A, Nair S, Tingley L, Day M, Mohamed I, Jenkins C, Godbole G. 2021. Phylogenomics and antimicrobial resistance of *Salmonella* Typhi and Paratyphi A, B and C in England, 2016-2019. Microb Genom 7:000633. https://doi.org/10.1099/mgen.0.000633
27. Kariuki S, Dyson ZA, Mbae C, Ngetich R, Kavai SM, Wairimu C, Anyona S, Gitau N, Onsare RS, Ongandi B, Duchene S, Ali M, Clemens JD, Holt KE, Dougan G. 2021. Multiple introductions of multidrug-resistant typhoid associated with acute infection and asymptomatic carriage, Kenya. Elife 10:e67852. https://doi.org/10.7554/eLife.67852
28. Gao Q, Liu Z, Xiang J, Zhang Y, Tong MX, Wang S, Zhang Y, Liu Q, Jiang B, Bi P. 2021. Impact of temperature and rainfall on typhoid/paratyphoid fever in Taizhou, China: effect estimation and vulnerable group identification. Am J Trop Med Hyg 106:532–542. https://doi.org/10.4269/ajtmh.20-1457
29. Saad NJ, Lynch VD, Antillón M, Yang C, Crump JA, Pitzer VE. 2018. Seasonal dynamics of typhoid and paratyphoid fever. Sci Rep 8:6870. https://doi.org/10.1038/s41598-018-25234-w
30. Phillips MT, Owers KA, Grenfell BT, Pitzer VE. 2020. Changes in historical typhoid transmission across 16 U.S. cities, 1889-1931: quantifying the

impact of investments in water and sewer infrastructures. PLoS Negl Trop Dis 14:e0008048. https://doi.org/10.1371/journal.pntd.0008048

31. Dong B-Q, Yang J, Wang X-Y, Gong J, von Seidlein L, Wang M-L, Lin M, Liao H-Z, Ochiai RL, Xu Z-Y, Jodar L, Clemens JD. 2010. Trends and disease burden of enteric fever in Guangxi province, China, 1994-2004. Bull World Health Organ 88:689–696. https://doi.org/10.2471/BLT.09.069310
32. Zhang H, Zhang X, Yan M, Pang B, Kan B, Xu H, Huang X. 2011. Genotyping of *Salmonella enterica* serovar Typhi strains isolated from 1959 to 2006 in China and analysis of genetic diversity by genomic microarray. Croat Med J 52:688–693. https://doi.org/10.3325/cmj.2011.52.688
33. Feng Y, Zou S, Chen H, Yu Y, Ruan Z. 2021. BacWGSTdb 2.0: a one-stop repository for bacterial whole-genome sequence typing and source tracking. Nucleic Acids Res 49:D644–D650. https://doi.org/10.1093/nar/gkaa821
34. Ruan Z, Yu Y, Feng Y. 2020. The global dissemination of bacterial infections necessitates the study of reverse genomic epidemiology. Brief Bioinform 21:741–750. https://doi.org/10.1093/bib/bbz010
35. Dyson ZA, Klemm EJ, Palmer S, Dougan G. 2019. Antibiotic resistance and typhoid. Clin Infect Dis 68:S165–S170. https://doi.org/10.1093/cid/ciy1111
36. Thanh Duy P, Thieu NTV, Nguyen Thi Nguyen T, Ngoc Dan Thanh H, Dongol S, Karkey A, Carey M, Basnyat B, Dougan G, Rabaa MA, Baker S. 2020. Gallbladder carriage generates genetic variation and genome degradation in *Salmonella* Typhi. PLoS Pathog 16:e1008998. https://doi.org/10.1371/journal.ppat.1008998
37. Prjibelski A, Antipov D, Meleshko D, Lapidus A, Korobeynikov A. 2020. Using SPAdes *de novo* assembler. Curr Protoc Bioinformatics 70:e102. https://doi.org/10.1002/cpbi.102
38. Zhang S, den Bakker HC, Li S, Chen J, Dinsmore BA, Lane C, Lauer AC, Fields PI, Deng X. 2019. SeqSero2: rapid and improved *Salmonella* serotype determination using whole-genome sequencing data. Appl Environ Microbiol 85:e01746-19. https://doi.org/10.1128/AEM.01746-19
39. Yoshida CE, Kruczkiewicz P, Laing CR, Lingohr EJ, Gannon VPJ, Nash JHE, Taboada EN. 2016. The *Salmonella in silico* typing resource (SISTR): an open web-accessible tool for rapidly typing and subtyping draft *Salmonella* genome assemblies. PLoS One 11:e0147101. https://doi.org/10.1371/journal.pone.0147101
40. Chen J, Ed-Dra A, Zhou H, Wu B, Zhang Y, Yue M. 2022. Antimicrobial resistance and genomic investigation of non-typhoidal *Salmonella* isolated from outpatients in Shaoxing city, China. Front Public Health 10:988317. https://doi.org/10.3389/fpubh.2022.988317
41. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, Parkhill J, Harris SR. 2015. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic Acids Res 43:e15. https://doi.org/10.1093/nar/gku1196
42. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. 2020. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol 37:2461. https://doi.org/10.1093/molbev/msaa131
43. Letunic I, Bork P. 2021. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res 49:W293–W296. https://doi.org/10.1093/nar/gkab301
44. Cheng L, Connor TR, Sirén J, Aanensen DM, Corander J. 2013. Hierarchical and spatially explicit clustering of DNA sequences with BAPS software. Mol Biol Evol 30:1224–1228. https://doi.org/10.1093/molbev/mst028
45. Rambaut A, Lam TT, Max Carvalho L, Pybus OG. 2016. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). Virus Evol 2:vew007. https://doi.org/10.1093/ve/vew007
46. Rieux A, Khatchikian CE. 2017. tipdatingbeast: an r package to assist the implementation of phylogenetic tip-dating tests using beast. Mol Ecol Resour 17:608–613. https://doi.org/10.1111/1755-0998.12603
47. Duchêne S, Holt KE, Weill F-X, Le Hello S, Hawkey J, Edwards DJ, Fourment M, Holmes EC. 2016. Genome-scale rates of evolutionary change in bacteria. Microb Genom 2:e000094. https://doi.org/10.1099/mgen.0.000094
48. Bouckaert R, Vaughan TG, Barido-Sottani J, Duchêne S, Fourment M, Gavryushkina A, Heled J, Jones G, Kühnert D, De Maio N, Matschiner M, Mendes FK, Müller NF, Ogilvie HA, du Plessis L, Popinga A, Rambaut A, Rasmussen D, Siveroni I, Suchard MA, Wu C-H, Xie D, Zhang C, Stadler T, Drummond AJ, Pertea M. 2019. BEAST 2.5: an advanced software platform for Bayesian evolutionary analysis. PLoS Comput Biol 15:e1006650. https://doi.org/10.1371/journal.pcbi.1006650
49. Bortolaia V, Kaas RS, Ruppe E, Roberts MC, Schwarz S, Cattoir V, Philippon A, Allesoe RL, Rebelo AR, Florensa AF, Fagelhauer L, Chakraborty T, Neumann B, Werner G, Bender JK, Stingl K, Nguyen M, Coppens J, Xavier BB, Malhotra-Kumar S, Westh H, Pinholt M, Anjum MF, Duggett NA, Kempf I, Nykäsenoja S, Olkkola S, Wieczorek K, Amaro A, Clemente L, Mossong J, Losch S, Ragimbeau C, Lund O, Aarestrup FM. 2020. ResFinder 4.0 for predictions of phenotypes from genotypes. J Antimicrob Chemother 75:3491–3500. https://doi.org/10.1093/jac/dkaa345
50. Carattoli A, Hasman H. 2020. PlasmidFinder and *in silico* pMLST: identification and typing of plasmid replicons in whole-genome sequencing (WGS). Methods Mol Biol 2075:285–294. https://doi.org/10.1007/978-1-4939-9877-7_20
51. Li Y, Teng L, Xu X, Li X, Peng X, Zhou X, Du J, Tang Y, Jiang Z, Wang Z, Jia C, Müller A, Kehrenberg C, Wang H, Wu B, Weill F-X, Yue M. 2022. A nontyphoidal *Salmonella* serovar domestication accompanying enhanced niche adaptation. EMBO Mol Med 14:e16366. https://doi.org/10.15252/emmm.202216366
52. Li Y, Ed-Dra A, Tang B, Kang X, Müller A, Kehrenberg C, Jia C, Pan H, Yang H, Yue M. 2022. Higher tolerance of predominant *Salmonella* serovars circulating in the antibiotic-free feed farms to environmental stresses. J Hazard Mater 438:129476. https://doi.org/10.1016/j.jhazmat.2022.129476
53. Kang X, Zhou X, Tang Y, Jiang Z, Chen J, Mohsin M, Yue M. 2022. Characterization of two-component system CitB family in *Salmonella* pullorum. Int J Mol Sci 23:10201. https://doi.org/10.3390/ijms231710201
54. Chen J, Zhou X, Tang Y, Jiang Z, Kang X, Wang J, Yue M. 2023. Characterization of two-component system CitB family in *Salmonella enterica* serovar gallinarum biovar gallinarum. Vet Microbiol 278:109659. https://doi.org/10.1016/j.vetmic.2023.109659
55. Achtman M, Zhou Z, Alikhan N-F, Tyne W, Parkhill J, Cormican M, Chiou C-S, Torpdahl M, Litrup E, Prendergast DM, Moore JE, Strain S, Kornschober C, Meinersmann R, Uesbeck A, Weill F-X, Coffey A, Andrews-Polymenis H, Curtiss Rd R, Fanning S. 2020. Genomic diversity of *Salmonella enterica* -the UoWUCC 10K genomes project. Wellcome Open Res 5:223. https://doi.org/10.12688/wellcomeopenres.16291.2

Genes, Brain and Behavior (2006) **5**: 209–221

Copyright © Blackwell Munksgaard 2005

Review

Early genomics of learning and memory: a review

S. Paratore†, E. Alessi‡, S. Coffa‡, A. Torrisi†, F. Mastrobuono† and S. Cavallaro†,*

†Institute of Neurological Sciences, Italian National Research Council, and ‡STMicroelectronics, Catania, Italy

**Corresponding author: S. Cavallaro, MD, PhD, Istituto di Scienze Neurologiche, CNR, Viale Regina Margherita 6, 95123 Catania, Italy. E-mail:s.cavallaro@isn.cnr.it*

The characterization of the molecular mechanisms whereby our brain codes, stores and retrieves memories remains a fundamental puzzle in neuroscience. Despite the knowledge that memory storage involves gene induction, the identification and characterization of the effector genes has remained elusive. The completion of the Human Genome Project and a variety of new technologies are revolutionizing the way these mechanisms can be explored. This review will examine how a genomic approach can be used to dissect and analyze the complex dynamic interactions involved in gene regulation during learning and memory. This innovative approach is providing information on a new class of genes associated with learning and memory in health and disease and is elucidating new molecular targets and pathways whose pharmacological modulation may allow new therapeutic approaches for improving cognition.

Keywords: Behavior, genomics, learning, memory, microarray

Received 15 February 2005, revised 29 April 2005, accepted for publication 29 April 2005

One of the most ambitious goals of modern neuroscience is to identify the mechanisms whereby our brain codes, stores and retrieves memories. Two general forms of memory can be classified by their duration: short-term memory (STM), which is rapidly formed and can outlast training for minutes or hours, and long-term memory (LTM), which lasts from hours to days, weeks or even years. STM involves post-translational modifications of preexisting molecules that alters the efficiency of synaptic transmission. In contrast, LTM can be blocked by inhibitors of transcription or translation indicating that it is dependent on *de novo* gene expression (Davis & Squire 1984; Stork & Welzl 1999). Proteins newly synthesized during memory consolidation may contribute to restructuring processes at the synapse and thereby alter the efficiency of synaptic transmission beyond the duration of STM. Revealing the dependence of LTM on protein synthesis, however, provides no information about the identity and specificity of the required proteins.

Because the quantity of a particular protein is often reflected by the abundance of its mRNA, a variety of methods has been used to describe only a limited number of differentially expressed genes during LTM (Stork & Welzl 1999). Successive screening therefore were needed to uncover how many and which genes are involved in memory and how do they interact functionally to effect memory storage. To survey the gene-based molecular mechanisms that underlie LTM, genome-wide expression analysis is being used in a variety of behavioral conditions (Cavallaro *et al.* 2001; Cavallaro *et al.* 2002; D'Agata & Cavallaro 2003; Dubnau *et al.* 2003; Leil *et al.* 2003; Luo *et al.* 2001; Rampon *et al.* 2000). This genomic approach requires the development of various laboratory protocols, as well as the development of database and software tools for efficient data collection and analysis. A basic understanding of these computational tools is therefore required for optimal experimental design and meaningful data analysis. For this reason, in the present review, we will outline the main procedures of microarray analysis (Fig. 1) before describing its application in the learning and memory field. For a more general description of microarray technology, the reader is referred to other reviews (Heller 2002; Hess *et al.* 2001; Noordewier & Warren 2001; Quackenbush 2001).

Expression profiling by DNA microarray technology

Basic principles

A DNA microarray is a grid of DNA spots, called probes, each containing a unique DNA sequence (Fig. 2). Spots contain either DNA oligomers or a longer DNA sequence designed to be complementary to a particular mRNA of interest. When a microarray is hybridized to fluorescence-tagged complementary DNAs or RNAs derived from messenger or total RNA, each spot is a target for the mRNA encoded by a gene. A laser can then excite the bound cDNAs or cRNAs and a scanner collects fluorescence intensities from each spot on the slide. The intensity of the fluorescence at each array element is proportional to the expression level of that gene in the sample. The choice of having oligomers or longer

doi: 10.1111/j.1601-183X.2005.00159.x

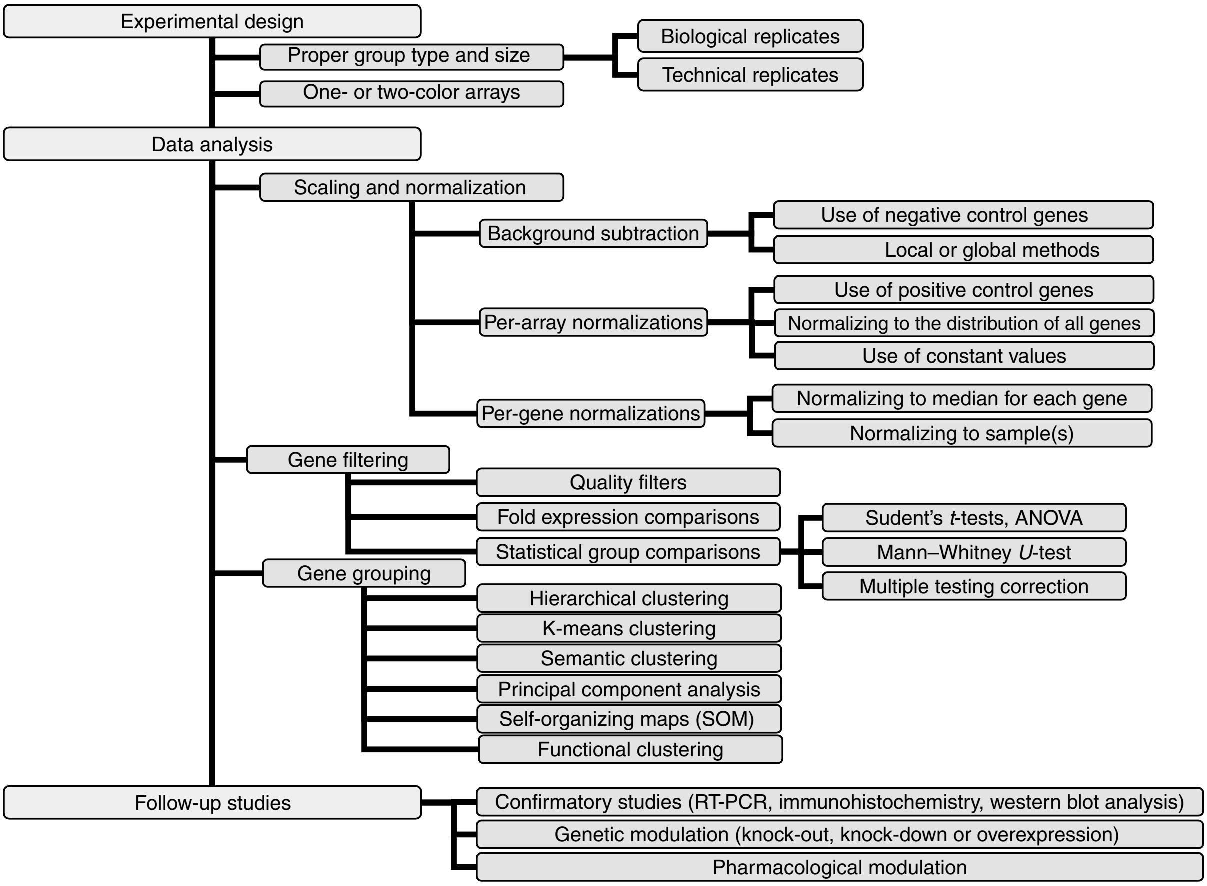

Experimental design
Proper group type and size
One- or two-color arrays
Biological replicates
Technical replicates
Data analysis
Scaling and normalization
Background subtraction
Use of negative control genes
Local or global methods
Per-array normalizations
Use of positive control genes
Normalizing to the distribution of all genes
Use of constant values
Per-gene normalizations
Normalizing to median for each gene
Normalizing to sample(s)
Gene filtering
Quality filters
Fold expression comparisons
Statistical group comparisons
Sudent's *t*-tests, ANOVA
Mann–Whitney *U*-test
Multiple testing correction
Gene grouping
Hierarchical clustering
K-means clustering
Semantic clustering
Principal component analysis
Self-organizing maps (SOM)
Functional clustering
Follow-up studies
Confirmatory studies (RT-PCR, immunohistochemistry, western blot analysis)
Genetic modulation (knock-out, knock-down or overexpression)
Pharmacological modulation

Figure 1: Experimental design, data analysis and follow-up experiments in DNA microarray experiments. DNA microarray is a very powerful technology, but it can be useless if the experimental design is not properly identified and executed with precision and accuracy. Because of variability across different samples, it is important to perform microarray experiments on replicate samples. These replicates can be either 'technical', where the same sample pair is repeated, or 'biological', where the experiment is repeated and an equivalent sample used. Sophisticated computational tools are then needed for the management and analysis of data that result from these experiments. Microarray data are commonly normalized to correct differences in intensity levels. Once normalized, a series of restrictions (gene filtering) can be applied to the data obtained. These restrictions include factors such as quality control, expression level constraints, sample-to-sample fold comparison and statistical group comparisons. More complex computational methods are needed to investigate and interpret microarray data and to group genes based on their expression levels or gene ontologies. Follow-up experiments can finally be used to confirm/modulate the expression/activity of candidate genes or their encoded proteins.

cDNA sequences yields two different microarray technologies: oligonucleotide and cDNA microarrays, respectively. The thing that makes microarrays the most promising technology for genome-wide expression analysis is the number of DNA probes that it is possible to place on a microarray. Already there are microarrays with probes for every gene in yeast, and others with over 40 000 human genes. This allows researchers to observe the response of whole genomes to various stimuli instead of one gene at a time.

Computational analysis of microarray data

Microarray analysis results in large amounts of data that are difficult to interpret without computational methods. The simplest analysis involves two samples, representing a test condition and a control condition and yields a list of paired expression values, one pair for each gene. As illustrated in Fig. 2, these pairs can be represented graphically by a scatter plot, with the values of sample one plotted on the x-axis and the values of sample two plotted on the y-axis. The resulting correlation plot provides a visual image of the relationship between the two expression profiles. In this plot, genes with similar expression levels in the two samples should have points on the identity line ($y = x$), and genes that are expressed differentially lie at some distance from this line. However, the problem is that microarrays do not measure expression levels directly, but rather intensity levels, as represented by the amount of phosphorescent dye that was recorded by a scanner. Many other factors, such as the overall mRNA concentration of the two samples, the saturation

1601183x, 2006, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/j.1601-183X.2005.00159.x, Wiley Online Library on [13/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

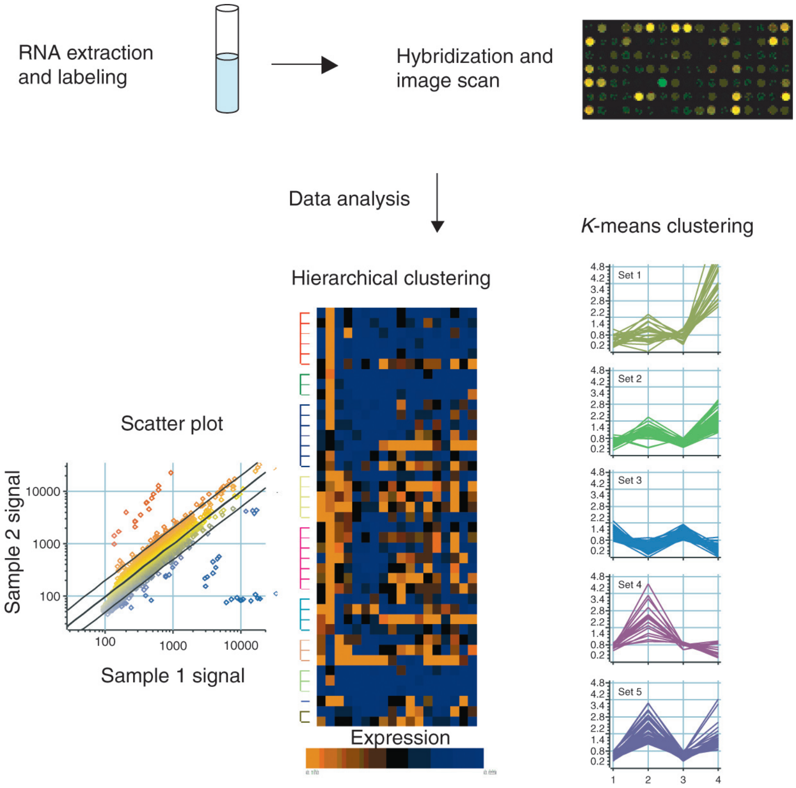


Figure 2: DNA microarray methodology. Total or messenger RNA is extracted, reverse transcribed, labeled with different fluorochromes (e.g. Cy3 and Cy5) and hybridized to microarrays. At the end of the hybridization, a laser scanner collects the image produced by the dye. Intensity values from each spot are calculated and then analyzed by specific software. Data can be represented graphically by a scatter plot, with the values of sample one plotted on the *x*-axis and the values of sample two plotted on the *y*-axis. Data obtained under different conditions (e.g. different time points) can be analyzed with different algorithms such as hierarchical or *k*-mean clustering.

effects in the hybridization or the quenching effect of the phosphorescent dyes can affect these intensity values. In order to correct these differences in intensity levels, the raw data can be normalized, for example, by using a normalization constant derived from housekeeping or spiked control genes.

Once normalized, a series of restrictions (or filters) can be applied to the data obtained. These restrictions include factors such as quality control, expression level constraints, sample-to-sample fold comparison and statistical group comparisons. The simplest way to identify interesting genes in DNA-microarray experiments is to search for those that are consistently either up- or downregulated. Relative differences in expression levels (fold changes) have been typically employed in group comparisons of gene expression. This approach, however, is somewhat arbitrary and inherently subject to high error rates because information on sample variance is not exploited. If array experiments are replicated to an extent that permits direct estimates of the variance of each individual transcript, parametric or non-parametric statistics can be applied. In these cases, however, a high number of false-positive results are expected by chance when one relies on the nominal *P*-value. For instance, when testing 10 000 transcripts we would expect to misidentify about 500 genes as significant ($P < 0.05$), even when there is no real difference in gene expression. Multiple testing corrections therefore are needed to adjust the individual *P*-value to account for this effect.

More complex computational methods are needed to monitor several gene expression profiles, such as those arising from time-course studies, and various clustering techniques have been applied to the identification of patterns in gene-expression data. Cluster analysis is a commonly used method to investigate and interpret gene expression data sets. By grouping together genes that have similar expression profiles, cluster analysis can be used for extraction of regulatory motifs, inference of functional annotation and classification of cell types or tissue samples.

Cluster analysis

The term *clustering* stands for a method that makes it possible to partition a set of objects (genes) into subgroups with similar features called clusters. These partitions have to satisfy the following features: *Homogeneity* in the cluster (the objects which belong to the same cluster have to be as similar as possible) and *Heterogeneity* among clusters (the objects which belong to different clusters have to be as different as possible).

A clustering method generally consists of two distinct components: a distance measure (or similarity coefficient) that indicates how similar two gene expression patterns are (or more generally, two clusters), and a clustering algorithm, which sorts the data and groups genes together on the basis of their separation in expression space.

Measure of distance or similarity coefficient

Many of the advanced analysis techniques are based upon measures of gene similarity. Similarity between genes is

1601183x, 2006, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/j.1601-183X.2005.00159.x, Wiley Online Library on [13/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

usually based on the correlation between the expression profiles of the genes. For expression data, we can solve the problem of 'similarity' mathematically by defining an 'expression vector' for each gene that represents its location in 'expression space'. In this way, expression data can be represented in *n*-dimensional expression space, where *n* is the number of experiments and where each gene-expression vector is represented as a single point in that data space.

In any clustering algorithm, the calculation of a 'distance' between any two objects is fundamental for placing them into groups. There are various methods for measuring distance, typically falling into two general classes: metric and semimetric (or similarity coefficient). Each of these takes two expression patterns and produces a number representing how similar the two genes are.

Detailed mathematical description of distance metrics used in clustering analysis can be retrieved in other reviews (Quackenbush 2001). The most common metric distance is the Euclidean distance. It simply is the geometric distance in the multidimensional space. The most commonly used semimetric distance measure in the analysis of gene expression data is the Pearson's correlation coefficient *r*.

Clustering algorithms

After providing means of measuring distance between genes, clustering algorithms sort the data and group genes together on the basis of their separation in expression space. Various clustering techniques have been applied to the identification of patterns in gene-expression data. Most cluster analysis techniques are hierarchical, the resultant classification has an increasing number of nested classes and the result resembles a phylogenetic classification. Non-hierarchical clustering techniques also exist, such as *k*-means clustering, which simply partition objects into different clusters without trying to specify the relationship between individual elements. Examples of hierarchical and *k*-means clustering are reproduced in Fig. 2.

Although cluster analysis techniques are extremely powerful, great care must be taken in applying this family of techniques. Eventhough the methods used are objective in the sense that the algorithms are well defined and reproducible, they are still subjective in the sense that selecting different algorithms, different normalizations, or different distance metrics, will place different objects into different clusters. Furthermore, clustering unrelated data would still produce clusters, although they might not be biologically meaningful.

Numerical, semantic and mixed clustering. Cluster analysis is a methodology to identify groups of genes that share common expression characteristics and behaviors. It has been frequently exploited in the analysis of genome-wide expression data, as the experimental observation that a set of genes is coexpressed implies that the genes share a biological function and are under common regulatory control.

Frequently, the clustering is used to group together genes considering only similar expression profiles, but it does not consider other well-known features of the gene properties. Actually, genes with a different profile expression could have similar functions as well and the classical clustering methodologies do not put it in evidence.

In order to extract knowledge from gene expression information, cluster analysis can be organized in three different approaches: numerical, semantic and mixed clustering.

The *numerical clustering* method is applied to the levels of gene expression. It tends grouping genes with a similar expression profile in the same clusters and makes sure that genes having different profiles with similar semantic features fall in different clusters. These considerations suggest that simple numerical clustering algorithms are inadequate to infer the genes' and proteins' role.

In order to discover more complex relationships among gene sequences the *semantic clustering* is used. The term *semantic clustering* indicates methods of clustering based on semantic characteristics, such as gene ontologies. When categorical domains are ordered, they can be turned in numerical values in order to transform those that are similar in near values. After that, methods of classical numerical clustering can be applied to that data set. When categorical domains are not ordered, however, this approach does not necessarily produce meaningful results. In this case, the k-modes algorithm can be used to remove this limitation.

Finally, more useful analyses can be performed using the *mixed* clustering. In this case, each gene can be represented from a vector in the n + m-dimensional space, where n is the number of levels of gene expression and m is the number of semantic features transformed in numerical values. Then, each gene-expression-semantic vector is represented as a single point in data space and any measure of distance can be adopted to calculate the distance between any two genes. The *mixed* clustering tends to group genes with similar expression profiles as well as genes with similar semantic features.

In order to perform *semantic* and *mixed* clustering we have developed new informatics applications (Fig. 3). Functional information is automatically retrieved by means of the software application GENELINK (Fig. 3A). This application is designed to retrieve genomics and proteomics information from external worldwide databases (NCBI GenBank and LocusLink). Before performing the semantic clustering, the functional annotations have to be turned into numerical values to transform features that are similar functionally in near values. For each gene ontology (GO) number, for example, the software application GENE ONTOLOGY SYSTEM (GOS) assigns a new GO number in order to identify the hierarchical relationships among GOs (Fig. 3b). The basic idea is that two related GOs must be coded with two closest GO numbers. After this renumbering process, methods of classical numerical clustering can be applied to the original data set (Fig. 3c) to perform numerical (Fig. 3d), semantic (Fig. 3e), or mixed (Fig. 3f) clustering. In this way, cluster analysis can effectively extract functional information from gene expression data.

1601183x, 2006, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/j.1601-183X.2005.00159.x, Wiley Online Library on [13/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

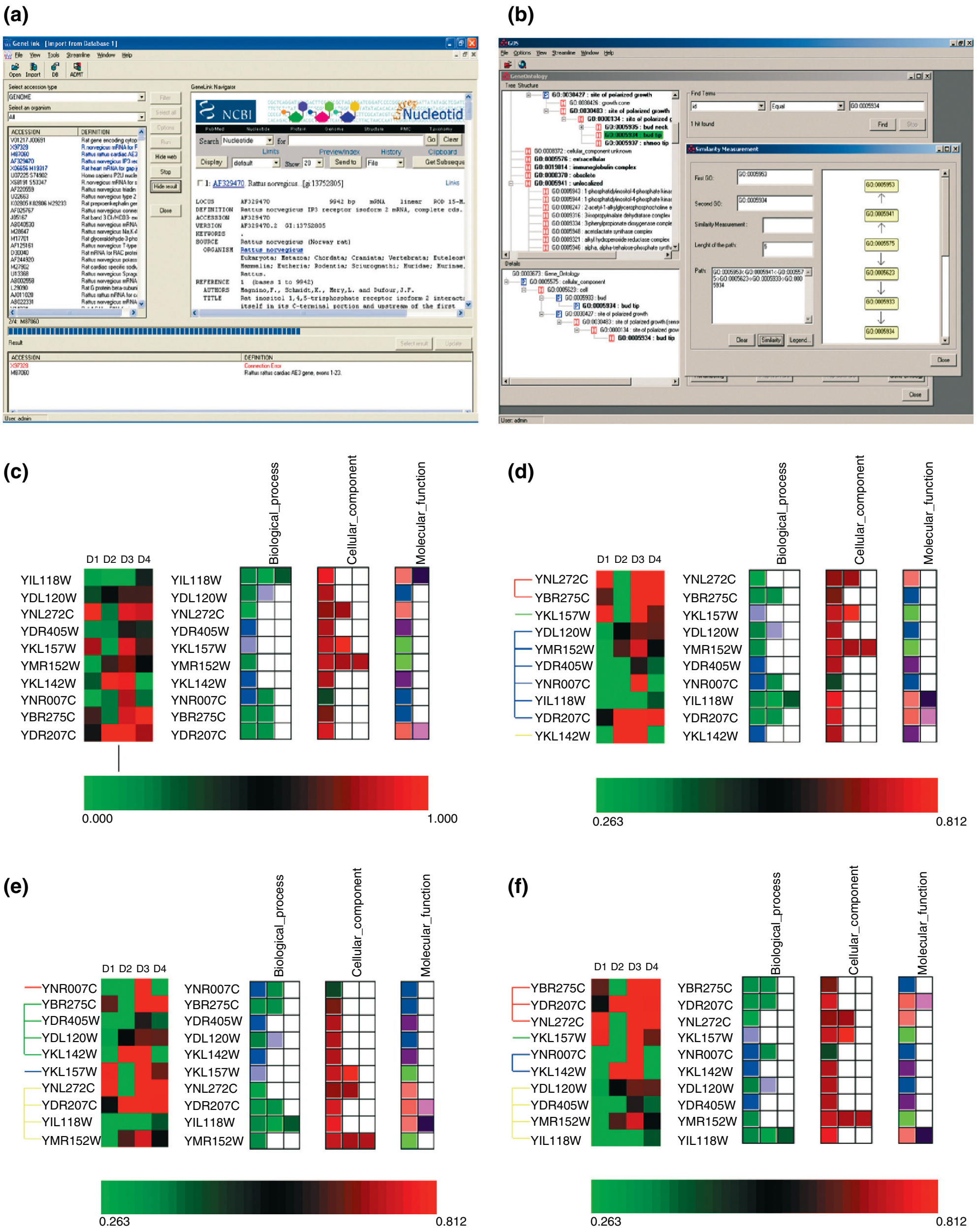


Figure 3: Numerical, semantic and mixed clustering. This term indicates a method of clustering genes based on their semantic characteristics. In this figure, gene ontologies (GOs) terms are used as examples of semantic features. Currently, three independent ontologies' information are accessible on the World Wide Web (http://www.geneontology.org): biological process (GO1), molecular function (GO2) and cellular component (GO3). We have developed an informatics application that we call GENELINK that is capable to automatically retrieve functional annotations from external databases (NCBI GenBank and LocusLink) (panel a). Before performing the semantic clustering these functional annotations have to be turned in numerical values in order to transform features that are similar functionally in near values. For each GO number, the software application GENE ONTOLOGY SYSTEM (GOS) assigns a new number according to a depth visit of the tree that represents the hierarchical relationships among GOs (Panel b). After this renumbering process, methods of classical numerical clustering can be applied to the normalized data set. Panels c–f show a gene expression matrix (left) and a GO map (right). The GO map is composed of three grids, one for each type of gene ontology: biological process, cellular component and molecular function. Each element of a grid is colored on the basis of the GO number obtained by GO renumbering. (some genes have more than one GO). Panel c represents unclustered gene expression data of 10 genes in four different conditions together with their GO map. Panel d shows clustered gene expression data of the 10 genes in four different conditions (GOs are not grouped). In panel e, gene ontologies are clustered (gene expression data are not grouped). Panel f shows a mixed (gene expression and GO data together) clustering (balanced clusters on both the universes can be seen).

1601183x, 2006, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/j.1601-183X.2005.00159.x, Wiley Online Library on [13/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

The use of a genomic approach for studying learning and memory

The following part of this review will focus on the use of a genomic approach to dissect and analyze gene-based mechanisms underlying learning and memory. We will highlight gene expression microarray analysis performed in different behavioral paradigms (eye-blink conditioning, water- and T-maze learning and passive avoidance conditioning). For space limitation, in this review we will try to give a broad view of the results obtained refraining from discussing each of the genes implicated by microarray analysis.

Eye-blink conditioning

To begin a comprehensive survey of the molecular mechanisms that underlie LTM, we used cDNA microarray technology to perform genome-wide expression analysis after classical conditioning of the rabbit's nictitating membrane response (NMR), a uniquely well-controlled associative learning paradigm (Fig. 4a) (Cavallaro *et al.* 2001). Classical conditioning of the rabbit NMR involves the presentation of an innocuous stimulus such as a tone followed by a noxious stimulus such as air puff to or electrical stimulation around the eye (Gormezano *et al.* 1962). Extensive lesion and recording data have implicated the cortex of the cerebellum and in particular, lobule HVI, in classical conditioning of the rabbit NMR (Berthier & Moore 1986; Gould & Steinmetz 1996; Gruart & Yeo 1995; Schreurs *et al.* 1991; Yeo *et al.* 1985). Although the hippocampus may not be necessary for NMR conditioning, recording data do show consistent eye blink-conditioning-specific hippocampal changes (Coulter *et al.* 1989; Sanchez-Andres & Alkon 1991). In addition, imaging studies have implicated both structures in human eye blink conditioning (Blaxton *et al.* 1996; Logan & Grafton 1995; Molchan *et al.* 1994; Schreurs *et al.* 1997).

Messenger RNA levels from cerebellar lobule HVI and hippocampus of unpaired and paired rabbits were simultaneously analyzed with high-density cDNA microarrays containing more than 8700 cDNAs (Cavallaro *et al.* 2001). When gene expression patterns were compared, mRNA levels of 79 and 17 genes differed more than twofold in lobule HVI and hippocampus, respectively (Figs 4b,c). These genes were operationally defined as 'memory related genes' (MRGs). Approximately 50% of the MRGs differentially expressed in the hippocampus were also differentially expressed in the HVI lobule, suggesting common mechanisms of memory storage in the two areas.

A majority of MRGs were downregulated, whereas only two genes that differed by a factor greater than 2 were upregulated in lobule HVI of paired animals (Fig. 4b). Because LTM can be blocked by transcription and protein synthesis inhibitors, most previous reports have focused on the identification of proteins whose expression is upregulated (Davis & Squire 1984). The preponderant reduction of gene expression during LTM therefore would not have been predicted and provides new and unexpected insights into the molecular mechanisms that underlie it. The specific role of the downregulation of MRGs following learning remains a matter of speculation. Downregulation of a gene may be the end point in a dynamic gene expression process that begins with upregulation during acquisition of the learned response. Alternatively, memory storage may require a balance of upregulation of some genes and downregulation of genes that exert inhibitory constraints on memory formation (Alberini *et al.* 1994). These latter genes might be termed memory suppressor genes (Abel & Kandel 1998).

A majority of the MRGs implicated have no currently recognized function and are not yet named. Complete nucleotide sequence determination, conceptual translation, expression monitoring and biochemical analysis are currently underway (D'Agata *et al.* 2003) and should provide a detailed functional understanding of these genes. Seventeen genes have significant similarity to known genes and can be grouped into different functional classes (Fig. 4D).

Our microarray analysis of eye blink-conditioned rabbits (Cavallaro *et al.* 2001) was the first reported in the literature to demonstrate the feasibility and utility of a cDNA microarray system as a means of dissecting the molecular mechanisms of associative memory. Further studies, however, were required at different time points and behavioral conditions to better understand the role of the implicated genes. To perform such studies we and others have moved to rat or mice, two animal species that are better suitable for genomic studies than rabbit in terms of sequenced genes and available microarrays.

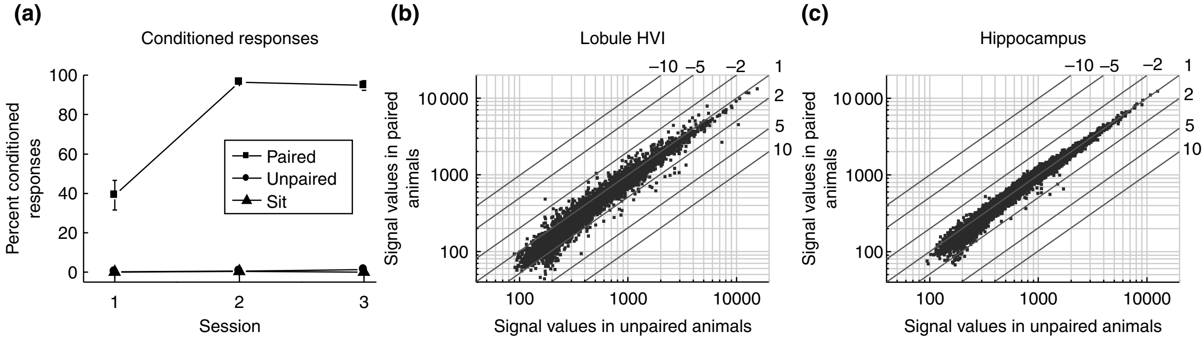
T-Maze learning

Microarrays have been used to analyze hippocampal gene expression in rats following training in a multiunit T-maze (Luo *et al.* 2001). In this study, the expression of 28 genes (18 known genes and 10 ESTs) was found to be increased in maze-trained animals compared with yoked control rats that were trained in a straight runway. Some of the known genes are involved in Ca^{2+} signaling, Ras activation, kinase cascades and extracellular matrix function. None of them, however, overlap with genes identified by microarray analysis in the other experimental paradigms examined in this review. Although the aversive foot shock was applied in equal duration and frequency to both the trained and control rats, changes in gene expression could be ascribed to other differences among the two groups, such as locomotor activity. In addition, because the animals were pretrained on day 1 and T-maze-trained on days 2 and 3, the time-dependent patterns of gene regulation during acquisition and consolidation of memory are unknown.

Water-maze learning

To detect learning related changes, microarray analysis has been used by two laboratories to characterize gene expression profiles in animals trained in the Morris water maze (Cavallaro *et al.* 2002; Leil *et al.* 2002; Leil *et al.* 2003). In this learning paradigm, a rodent learns to locate a submerged island in a large pool by

1601183x, 2006, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/j.1601-183X.2005.00159.x, Wiley Online Library on [13/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License



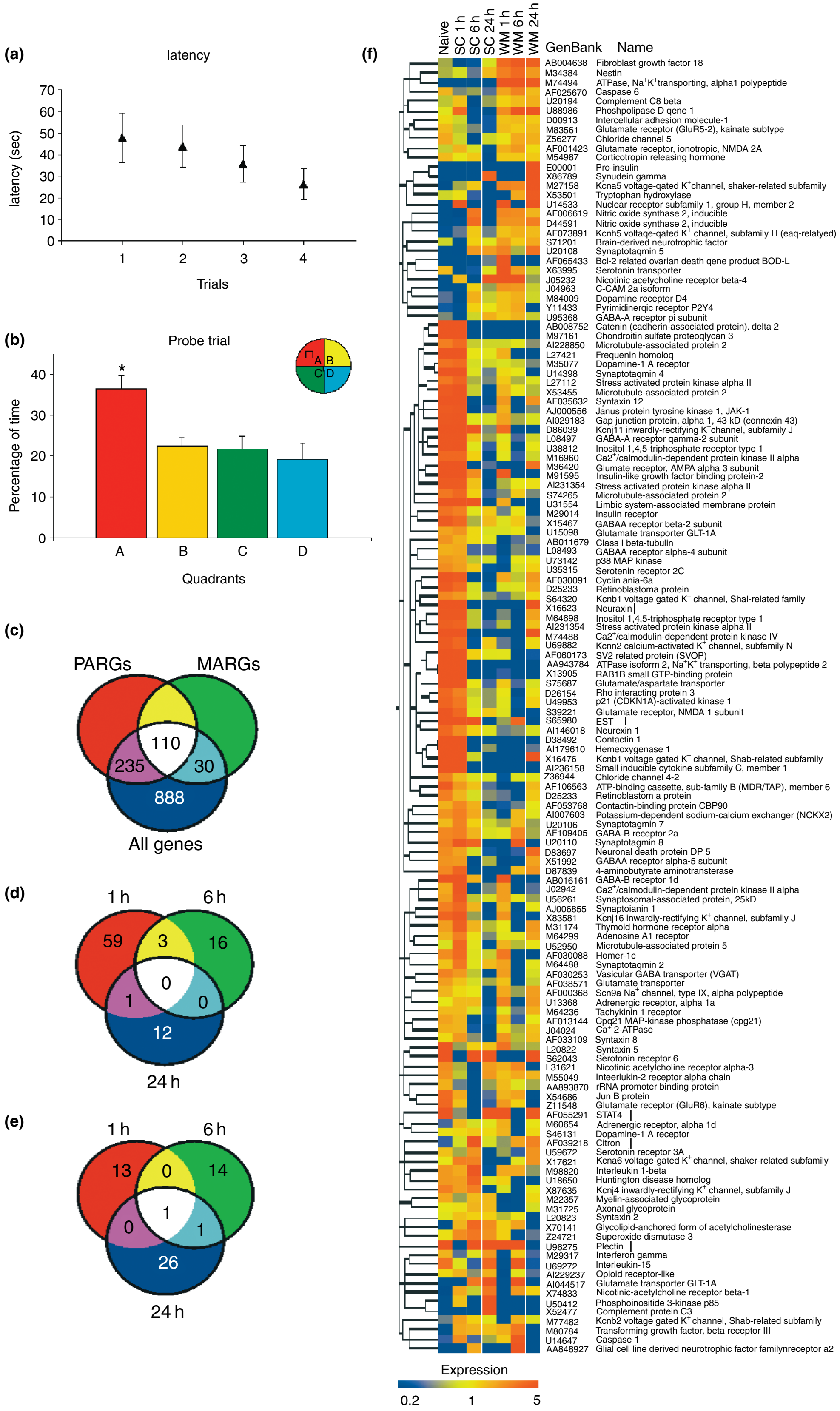
(d) Genes with a known function differentially expressed following NMR

Functional class	Fold change		Gene name	Accession number
	Hippocampus	Cerebellar lobule HVI		
Signal transduction		2.7	Insulin-like growth factor I	W10072.1
		2.3	Growth differentiation factor 9	AA422377.1
		2.3	Liprin-beta 2	W12503.1
		2.3	LAR receptor-linked tyrosine phosphatase	W16354.1
		2.2	Phocein	W50771.1
		2.2	Fibrinogen/angiopoietin-related protein	W13905.1
Protein modification		3.2	ESTs, Highly similar to CD156	AA470303.1
	2.3	2.4	F-box protein FBX8	AA413956.1
		2.3	ESTs, Higly similar to hippostasin	W13212.1
DNA transcription regulation	2.1	2.7	Bach protein 2	AI604485.1
		2.6	ESTs, Highly similar to WBSCR11	AA467436.1
		2.5	Pirin	W08720.1
	2.3		ESTs, Similar to C/EBP	AA473938.1
		2.3	DRG11	AA387552.1
		2.3	DCoH/PCD	W14332.1
		2.1	7SK	AA058055.1
		2.1	Orphan receptor TR2	AA163245.1

Figure 4: Microarray analyses of eye blink conditioned rabbits. (a) Mean percent conditioned responses in paired, unpaired and sit-control rabbits as a function of three training sessions. To relate changes in gene expression to a learning task, we used pairings of a tone and periorbital electrical stimulation in a standard delay-conditioning procedure, training rabbits to asymptotic levels of conditioning over 3 consecutive days. Paired rabbits ($n = 12$) acquired conditioned responses to the tone and reached a mean terminal level of 94.7% conditioned responses, whereas the unpaired control rabbits ($n = 12$) responded to the tone at mean levels of less than 1.3% across the 3 days of stimulus presentations and sit-control rabbits ($n = 5$) had spontaneous blink rates of less than 1% ($P < 0.001$). Without further training or testing rabbits show a level of 80% conditioned responses as long as 1 month after the 3 days of the stimulus pairings used in the present experiments. Consequently, harvesting cerebellar and hippocampal tissue 24 h after 3 days of pairings ensured that rabbits were still at an asymptotic level of conditioning. (b–c) Scatter plot of gene expression levels for paired and unpaired animals in (b) cerebellar lobule HVI and (c) hippocampus. Messenger RNA levels from cerebellar lobule HVI and hippocampus of unpaired and paired rabbits ($n = 7$ per group) were simultaneously analyzed with high-density cDNA microarrays containing more than 8700 cDNA mouse clones with a length of 500–5000 bp and with averages in the 1-kb region. The estimated percentage of homology between mouse clones and rabbit genes is 88.98 ± 3.7 (mean ± SD). The cross species similarity and a complete list of the differentially expressed genes are available online at http://web.tiscali.it/sebastiano_cavallaro. (D) Differentially expressed genes with a known function are ordered into functional groups.

creating a spatial map using extra-pool cues. Leil *et al.* (2002) used cDNA-microarrays containing approximately 9000 clones to detect hippocampal gene expression changes between F1 hybrid mouse strains that perform well on the Morris water maze and inbred strains that perform poorly. Although this study was performed in a brain region intimately involved in spatial learning, genes differentially expressed in mouse strains may subserve other behavioral processes or functions. Indeed, in a later study, the same authors (Leil *et al.* 2003) used microarray analysis to characterize the differential expression of genes ($n = 3$) in the hippocampus of F1 hybrid mice after 2 days of water-maze training. Although mouse strains used were the same, no overlap was found between the genes revealed in the two studies. In addition, no overlap was found between genes differentially expressed in water-maze training (Leil *et al.* 2003) and those performed in the other behavioral paradigms examined in this review. This is probably due to a number of factors, including different genes on the arrays, different species or strains employed, brain regions and time point studied. In addition, because mice are very reactive to placement in

1601183x, 2006, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/j.1601-183X.2005.00159.x, Wiley Online Library on [13/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License



1601183x, 2006, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/j.1601-183X.2005.00159.x, Wiley Online Library on [13/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Figure 5: Water-maze learning. (a) Escape latencies of rats swimming to a submerged platform in the water maze during four consecutive trials. In order to reduce stress in the experimental day, the first day was dedicated to swimming training in the absence of an island. Each rat was placed in the pool for 2 min and was returned to its home cage. In the next day, half of the rats were placed again in the pool for a 2.5-min swimming session and were used as swimming controls. The other half were given four consecutive trials to locate the platform, each trial lasting up to 2 min. Rats were required to spend 30 seconds of an intertrial interval on the platform. The rats' escape latency was measured using a HVS2020 video tracking system (HVS Image Ltd, Hampton, UK). (b) Probe trial. To verify that the trained rats in fact learned the spatial location of the island, a group of six rats was trained to find the island and tested 24 h later on a quadrant analysis test. The trained rats swam significantly longer in the quadrant (red) where the island was located. (c–e) Venn diagrams of differentially expressed hippocampal genes. Hippocampal gene expression profiles in naïve, SC and water-maze trained rats were measured using microarrays containing 1263 genes relevant to neurobiology (Affymetrix GeneChip Rat Neurobiology U34 array). Genes differentially expressed in naïve and swimming control animals 1, 6 and 24 h after training were operationally defined as 'physical activity related genes' (PARGs), whereas genes differentially expressed in water-maze-trained animals compared with swimming controls were operationally defined as 'memory related genes' (MRGs) (c). Among these, 55 genes were upregulated (d), whereas 91 genes were downregulated (e) in at least one of three time points examined. (f) Hierarchical clustering of MRGs. A hierarchical clustering algorithm (Pearson correlation, separation ratio 0.5, minimum distance 0.001) was used to order MRGs in a dendrogram in which the pattern and length of the branches reflects the relatedness of the samples. Data are presented in a matrix format: each row represents a single gene and each column an experimental condition. The averaged normalized intensity from two replicates is represented by the color of the corresponding cell in the matrix. Blue, yellow and red cells, respectively, represent transcript levels below, equal or above the median abundance across all conditions. Color intensity reflects the magnitude of the deviation from the median (see scale at the bottom). A complete list of the differentially expressed genes is available online at http://web.tiscali.it/sebastiano_cavallaro.

water, gene expression differences may be due to stress responses rather than learning and memory.

To analyze the time-dependent patterns of gene regulation during water-maze training, we measured hippocampal gene expression profiles in naïve, swimming control and water-maze trained rats (Fig. 5a,b), using microarrays containing more than 1200 genes relevant to neurobiology (Cavallaro *et al*. 2002). When gene expression profiles in naïve and swimming control animals 1, 6 and 24 h after swimming sessions were compared, 345 genes were found differentially expressed more than twofold in at least two of the four conditions (Fig. 5c). These genes, operationally defined as 'physical activity related genes' (PARGs) indicate that physical activity and mild stress associated with behavioral training has a significant impact on hippocampal gene expression.

When gene expression levels in swimming control animals were compared with water-maze trained animals 1, 6, or 24 h after training, 140 MRGs were found (Fig. 5c). The majority of these MRGs (110 of 140) were also PARGs, i.e. influenced by physical activity. Among MRGs, 55 genes were upregulated in the hippocampus of water-maze-trained animals, whereas 91 genes were downregulated (Figs 5d,e).

Most of the MRGs, those differentially expressed between the swimming and spatial learning animal groups, were also affected during swimming alone but with entirely different temporal patterns of expression (Fig. 5f). Although learning and physical activity involve common groups of genes, the behavior of learning and memory can be distinguished from unique patterns of gene expression across time.

All of the MRGs identified during water-maze learning have a recognized function and can be classified into six major groups based on their translated product: (i) cell signaling (ii) synaptic proteins (iii) cell–cell interaction and cytoskeletal proteins (iv) apoptosis (v) enzymes and (vi) transcription or translation regulation. Some of these genes have been previously related to synaptic plasticity, memory, or cognitive disorders. For a complete description of the MRGs implicated by microarray technology, the reader is referred to our previous study (Cavallaro *et al*. 2002). In the following paragraph we will discuss only one of the MRGs, FGF-18, which has been further tested for its memory regulatory function.

FGF-18 is a novel member of the FGF family, which was shown to stimulate neurite outgrowth (Ohbayashi *et al*. 1998). Although the function of this peptide is still unknown, the

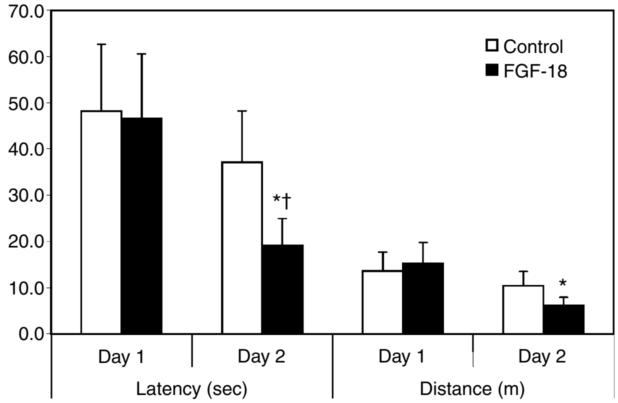
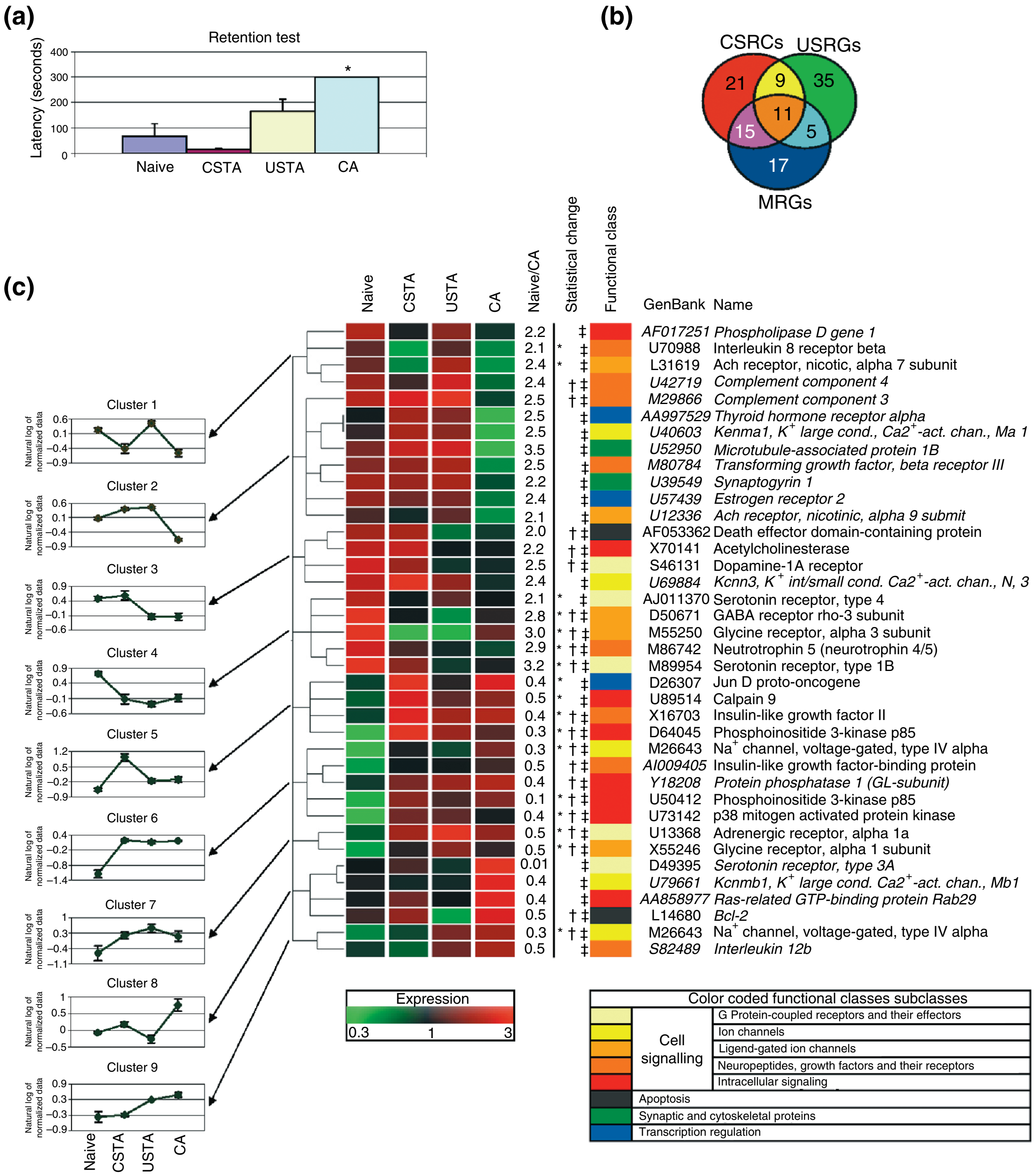


Figure 6: Effects of a single exogenous administration of FGF-18 on water-maze learning. Thirteen male Wistar rats (250–300 g) were implanted strereotaxically with stainless steel guide cannulae in the right and left lateral ventricles. On day 1, 1 week after surgery, animals were subjected to a 2-min swimming training session. Then, a water-maze training session was performed on days 2 and 3 and consisted in finding a submerged platform to escape from the water. Two trials were given to the animal for each session. The escape latency and distance to find the platform were monitored as described above. Ten minutes after the second trial on day 2, an intracerebroventricular administration of drug or vehicle was performed in both lateral ventricles. Six animals received 0.94 pmoles of FGF-18 and the other seven received a control injection of vehicle (saline). *day 1 vs. day 2, $P < 0.05$; †control vs. FGF-18, $P < 0.05$.

1601183x, 2006, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/j.1601-183X.2005.00159.x, Wiley Online Library on [13/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

other members of its family are important signaling molecules in several inductive and patterning processes and act as brain organizer-derived signals during formation of the early vertebrate nervous system. Water-maze training but not physical activity induced the expression of FGF-18. To explore the effect of FGF-18 in spatial learning, we tested the effects of a single exogenous dose of FGF-18. Rats were trained in a Morris water maze for two trials and then injected intracerebroventricularly with 0.94 pmoles of FGF-18 or vehicle. As shown in Fig. 6, FGF-18 treatment improved spatial learning behavior by inducing a 49% reduction in the escape latency but no significant changes in motor activity.

The data obtained in the hippocampus of water-maze-trained rats (Cavallaro *et al.* 2002) represent the first



1601183x, 2006, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/j.1601-183X.2005.00159.x, Wiley Online Library on [13/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Figure 7: Passive avoidance learning. (a) Passive avoidance retention test. Conditioned animals (CA) were trained to avoid moving from the lighted to the darkened section of a conditioning chamber by delivering a foot shock when they entered the darkened section. Control rats included untrained (naïve) animals, and animals exposed to the conditioned (CSTA) or the unconditioned (USTA) stimulus. Twenty-four hours after the training trial, half of the animals ($n = 4$ per group) performed the retention test to verify that the trained rats in fact learned the passive avoidance task. The animals were placed in the safe compartment with the door closed. After 2 min of acclimation the light turn-on, the door opened and the animal was allowed to enter the dark compartment. The latency to enter the dark compartment was recorded and used as the measure of retention. The rats avoiding the dark compartment for over 300 seconds were considered to have a memory of the training experience. During the retention trial, CA had a longer mean step-through latency than naïve, CSTA and USTA (*$P < 0.001$). (b) Venn diagrams of differentially expressed hippocampal genes. Hippocampal gene expression profiles in CA, USTA, CSTA and naïve animals were measured 6 h after training using microarrays containing 1263 genes relevant to neurobiology (Affymetrix GeneChip Rat Neurobiology U34 array). Genes differentially expressed between naïve and CSTA were defined as 'conditioned stimulus related genes' (CSRGs); genes differentially expressed between naïve and USTA were defined as 'unconditioned stimulus related genes' (USRGs); genes differentially expressed between naïve and CA were defined as 'memory related genes' (MRGs). (c) Hierarchical clustering of MRGs. A hierarchical clustering algorithm (Pearson correlation, separation ratio 0.2, minimum distance 0.001) was used to order MRGs in a dendrogram in which the pattern and length of the branches reflects the relatedness of the samples. Data are presented in a matrix format: each row represents a single gene and each column an experimental condition. The averaged normalized intensity from four replicates is represented by the color of the corresponding cell in the matrix. Green, black and red cells, respectively, represent transcript levels below, equal or above the median abundance across all conditions. Color intensity reflects the magnitude of the deviation from the median (see scale at the bottom). The graphs on the left of the dendrogram represent the averaged Natural Log of normalized data ± SEM of the genes in nine major clusters. Gene expression ratio between naïve and CA together with statistical significant changes ($P < 0.05$; *naïve vs. CSTA, †naïve vs. USTA and ‡naïve vs. CA) are shown on the right of the matrix. Functional classification of MRGs is represented in a column on the right of the figure where each functional classes or subclasses are color coded. The name and GenBank accession number of MRGs uniquely regulated in CA are indicated in italic, whereas MRGs previously found to be differentially expressed in the hippocampus of water-maze-trained rats (Cavallaro *et al.* 2002) are indicated in bold. A complete list of the differentially expressed genes is available online at http://web.tiscali.it/sebastiano_cavallaro.

temporal gene expression comparison reported in the long term retention of learning and memory and further demonstrated the utility of a genomic approach as a means of dissecting the molecular basis of associative memory. This approach provides information on the gene expression changes that occur during physical activity, stress, learning and memory, allowing the identification of molecular targets and pathways whose modulation may generate new therapeutic approaches for facilitating learning and memory.

Passive avoidance learning

We have recently extended our genome-wide screenings to an additional behavioral animal model, a step-through passive avoidance test, known to require hippocampus-dependent learning and depend upon transcription (Stubley-Weatherly *et al.* 1996). In these experiments (D'Agata & Cavallaro 2003), conditioned animals (CA) were trained to avoid moving from the lighted to the darkened section of a conditioning chamber by delivering a foot shock when they entered the darkened section. Control rats included untrained (naïve) animals, and animals exposed to the unconditioned (USTA) or the conditioned (CSTA) stimulus. To verify that the trained rats in fact learned the passive avoidance task, learning was assessed in a comparable group of animals by evaluating the latency of step-through in a retention test. Twenty-four hours after the one-trial training period, only CA learned to associate stepping through the darkened chamber with the foot shock (Fig. 7a).

Hippocampal gene expression profiles in CA, USTA, CSTA and naïve animals were measured 6 h after training using microarrays containing 1263 genes relevant to neurobiology (D'Agata & Cavallaro 2003). When gene expression profiles of naïve animals were compared with those of CSTA or USTA, 46 and 60 genes, respectively, were found differentially expressed (Fig. 7b). These genes further demonstrate that physical activity and mild stress associated with behavioral training have a significant impact on hippocampal gene expression.

When gene expression levels in naïve animals were compared with CA, 38 MRGs were found (Fig. 7b). Among these, 21 genes were downregulated and 17 genes were upregulated. Some of these MRGs (21/38) were also differentially expressed in CSTA (16) and USTA (16) (Fig. 7b).

A hierarchical clustering method was used to group MRGs on the basis of similarity in their expression patterns (Fig. 7c). The most evident traits of the clustered data were that MRGs showed entirely different patterns of expression in CA vs. CSTA or USTA. Genes segregating into nine major branches of the dendrogram were assigned to nine clusters (Fig. 7c). Clusters 1–4 represent those genes, which were downregulated, whereas clusters 5–9 include those, which were upregulated in CA. Some of the MRGs, those differentially expressed between naïve and CA, were also affected by exposing the rats to the conditioned or the unconditioned stimulus alone, whereas others were uniquely induced when the two were associated and the animals were conditioned (Fig. 7c, clusters 2 and 8). Expression changes of MRGs in CSTA or USTA had different magnitudes or more often opposite trends than CA (Fig. 7c, clusters 1, 2, 3, 5, 8 and 9). As we have previously observed in water-maze-trained animal, learning, physical activity and mild stress associated with behavioral training involve common groups of genes. Their behavior in learning and memory, however, could be

1601183x, 2006, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/j.1601-183X.2005.00159.x, Wiley Online Library on [13/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

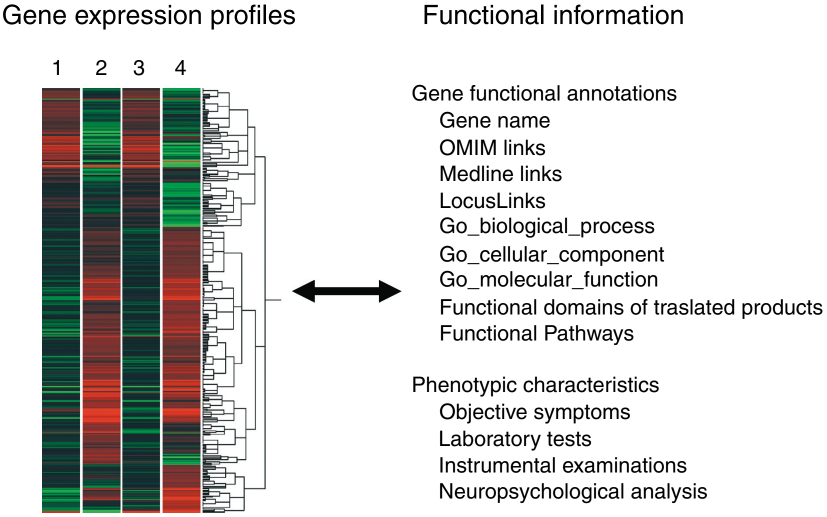


Figure 8: Correlations between gene expression profiles and functional information. Complex correlations between gene expression profiles and functional information are needed to unravel the role of genes in the pathophysiology of learning and memory. Examples of functional information are listed on the right side of the figure and include gene annotations and phenotypic characteristics (any identifiable or observable structural or functional characteristic of an organism).

distinguished from unique patterns of gene expression as shown in the clustered data.

All of the MRGs identified have a recognized function and can be classified into different functional classes based on their translated product (Fig. 7c). Some of these genes have been previously related to synaptic plasticity, memory or cognitive disorders. Six of 38 MRGs found in the hippocampus of rats after passive avoidance training (Fig. 7c, shown in bold) were also differentially expressed in the same brain area following water-maze learning (Fig. 5f) suggesting common mechanisms of memory storage in different behavioral paradigms. For a complete description of the MRGs implicated by microarray technology during passive avoidance conditioning, the reader is referred to our previous study (D'Agata & Cavallaro 2003).

Conclusions

The characterization of expression patterns associated to the long-term retention of learning and memory in health (Cavallaro *et al.* 2001; 2002; D'Agata & Cavallaro 2003; Dubnau *et al.* 2003; Luo *et al.* 2001; Leil *et al.* 2003; Rampon *et al.* 2000) and disease (Blalock *et al.* 2003; D'Agata *et al.* 2002; Hata *et al.* 2001; Ho *et al.* 2001; Leil *et al.* 2002; Loring *et al.* 2001; Pasinetti & Ho 2001; Pasinetti 2001; Tudor *et al.* 2002; Yao *et al.* 2003) conditions has just started. Gene expression profiles unlock virtually unexplored frontiers, and we will learn as we explore them.

These 'early' studies are limited by the experimental design (animal strain, behavioral condition and time), technology (microarray platforms and number of genes) and computational analysis (normalization, filtering and statistical analysis) used. The value of these experiments will progressively increase as more is learned about the function of each gene and when software applications, like that we presented in this paper, will enable us to identify complex correlations existing between the genomic profiles obtained by microarray experiments and functional information (Fig. 8).

Although sure to be just the tip of the iceberg, the results already obtained point toward genes or sets of genes that may play critical roles in learning and memory. The discovery of these genes represents the key to developing novel and efficacious therapies to improve learning and memory, under normal conditions as well as in disorders that affect cognitive functioning, such as Alzheimer's disease.

References

Abel, T. & Kandel, E. (1998) Positive and negative regulatory mechanisms that mediate long-term memory storage. *Brain Res Brain Res Rev* **26**, 360–378.

Alberini, C.M., Ghirardi, M., Metz, R. & Kandel, E.R. (1994) C/EBP is an immediate-early gene required for the consolidation of long-term facilitation in Aplysia. *Cell* **76**, 1099–1114.

Berthier, N.E. & Moore, J.W. (1986) Cerebellar Purkinje cell activity related to the classically conditioned nictitating membrane response. *Exp Brain Res* **63**, 341–350.

Blalock, E.M., Chen, K.C., Sharrow, K., Herman, J.P., Porter, N.M., Foster, T.C. & Landfield, P.W. (2003) Gene microarrays in hippocampal aging: statistical profiling identifies novel processes correlated with cognitive impairment. *J Neurosci* **23**, 3807–3819.

Blaxton, T.A., Bookheimer, S.Y., Zeffiro, T.A., Figlozzi, C.M., Gaillard, W.D. & Theodore, W.H. (1996) Functional mapping of human memory using PET: comparisons of conceptual and perceptual tasks. *Can J Exp Psychol* **50**, 42–56.

Cavallaro, S., D'Agata, V., Manickam, P., Dufour, F. & Alkon, D.L. (2002) Memory-specific temporal profiles of gene expression in the hippocampus. *Proc Natl Acad Sci USA* **99**, 16279–16284.

Cavallaro, S., Schreurs, B.G., Zhao, W., D'Agata, V. & Alkon, D.L. (2001) Gene expression profiles during long-term memory consolidation. *Eur J Neurosci* **13**, 1809–1815.

Coulter, D.A., Lo Turco, J.J., Kubota, M., Disterhoft, J.F., Moore, J.W. & Alkon, D.L. (1989) Classical conditioning reduces amplitude and duration of calcium-dependent afterhyperpolarization in rabbit hippocampal pyramidal cells. *J Neurophysiol* **61**, 971–981.

1601183x, 2006, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/j.1601-183X.2005.00159.x, Wiley Online Library on [13/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

D'Agata, V. & Cavallaro, S. (2003) Hippocampal gene expression profiles in passive avoidance conditioning. *Eur J Neurosci* **18**, 2835–2841.

D'Agata, V., Schreurs, B.G., Pascale, A., Zohar, O. & Cavallaro, S. (2003) Down regulation of cerebellar memory related gene-1 following classical conditioning. *Genes Brain Behav* **2**, 231–237.

D'Agata, V., Warren, S.T., Zhao, W., Torre, E.R., Alkon, D.L. & Cavallaro, S. (2002) Gene expression profiles in a transgenic animal model of fragile X syndrome. *Neurobiol Dis* **10**, 211–218.

Davis, H.P. & Squire, L.R. (1984) Protein synthesis and memory: a review. *Psychol Bull* **96**, 518–559.

Dubnau, J., Chiang, A.S., Grady, L., Barditch, J., Gossweiler, S., McNeil, J., Smith, P., Buldoc, F., Scott, R., Certa, U., Broger, C. & Tully, T. (2003) The staufen/pumilio pathway is involved in Drosophila long-term memory. *Curr Biol* **13**, 286–296.

Gormezano, I., Schneiderman, N., Deaux, E.G. & Fuentes, I. (1962) Nictitating membrane: Classical conditioning and extinction in the albino rabbit. *Science* **138**, 33–34.

Gould, T.J. & Steinmetz, J.E. (1996) Changes in rabbit cerebellar cortical and interpositus nucleus activity during acquisition, extinction, and backward classical eyelid conditioning. *Neurobiol Learn Mem* **65**, 17–34.

Gruart, A. & Yeo, C.H. (1995) Cerebellar cortex and eyeblink conditioning: bilateral regulation of conditioned responses. *Exp Brain Res* **104**, 431–448.

Hata, R., Masumura, M., Akatsu, H., Li, F., Fujita, H., Nagai, Y., Yamamoto, T., Okada, H., Kosaka, K., Sakanaka, M. & Sawada, T. (2001) Up-regulation of calcineurin Abeta mRNA in the Alzheimer's disease brain: assessment by cDNA microarray. *Biochem Biophys Res Commun* **284**, 310–316.

Heller, M.J. (2002) DNA MICROARRAY TECHNOLOGY: Devices, Systems, and Applications. *Annu Rev Biomed Eng* **4**, 129–153.

Hess, K.R., Zhang, W., Baggerly, K.A., Stivers, D.N. & Coombes, K.R. (2001) Microarrays: handling the deluge of data and extracting reliable information. *Trends Biotechnol* **19**, 463–468.

Ho, L., Guo, Y., Spielman, L., Petrescu, O., Haroutunian, V., Purohit, D., Czernik, A., Yemul, S., Aisen, P.S., Mohs, R. & Pasinetti, G.M. (2001) Altered expression of a-type but not b-type synapsin isoform in the brain of patients at high risk for Alzheimer's disease assessed by DNA microarray technique. *Neurosci Lett* **298**, 191–194.

Leil, T.A., Ossadtchi, A., Cortes, J.S., Leahy, R.M. & Smith, D.J. (2002) Finding new candidate genes for learning and memory. *J Neurosci Res* **68**, 127–137.

Leil, T.A., Ossadtchi, A., Nichols, T.E., Leahy, R.M. & Smith, D.J. (2003) Genes regulated by learning in the hippocampus. *J Neurosci Res* **71**, 763–768.

Logan, C.G. & Grafton, S.T. (1995) Functional anatomy of human eyeblink conditioning determined with regional cerebral glucose metabolism and positron-emission tomography. *Proc Natl Acad Sci U S A* **92**, 7500–7504.

Loring, J.F., Wen, X., Lee, J.M., Seilhamer, J. & Somogyi, R. (2001) A gene expression profile of Alzheimer's disease. *DNA Cell Biol* **20**, 683–695.

Luo, Y., Long, J.M., Spangler, E.L., Longo, D.L., Ingram, D.K. & Weng, N.P. (2001) Identification of maze learning-associated genes in rat hippocampus by cDNA microarray. *J Mol Neurosci* **17**, 397–404.

Molchan, S.E., Sunderland, T., McIntosh, A.R., Herscovitch, P. & Schreurs, B.G. (1994) A functional anatomical study of associative learning in humans. *Proc Natl Acad Sci U S A* **91**, 8122–8126.

Noordewier, M.O. & Warren, P.V. (2001) Gene expression microarrays and the integration of biological knowledge. *Trends Biotechnol* **19**, 412–415.

Ohbayashi, N., Hoshikawa, M., Kimura, S., Yamasaki, M., Fukui, S. & Itoh, N. (1998) Structure and expression of the mRNA encoding a novel fibroblast growth factor, FGF-18. *J Biol Chem* **273**, 18161–18164.

Pasinetti, G.M. (2001) Use of cDNA microarray in the search for molecular markers involved in the onset of Alzheimer's disease dementia. *J Neurosci Res* **65**, 471–476.

Pasinetti, G.M. & Ho, L. (2001) From cDNA microarrays to high-throughput proteomics. Implications in the search for preventive initiatives to slow the clinical progression of Alzheimer's disease dementia. *Restor Neurol Neurosci* **18**, 137–142.

Quackenbush, J. (2001) Computational analysis of microarray data. *Nat Rev Genet* **2**, 418–427.

Rampon, C., Jiang, C.H., Dong, H., Tang, Y.P., Lockhart, D.J., Schultz, P.G., Tsien, J.Z. & Hu, Y. (2000) Effects of environmental enrichment on gene expression in the brain. *Proc Natl Acad Sci U S A* **97**, 12880–12884.

Sanchez-Andres, J.V. & Alkon, D.L. (1991) Voltage-clamp analysis of the effects of classical conditioning on the hippocampus. *J Neurophysiol* **65**, 796–807.

Schreurs, B.G., McIntosh, A.R., Bahro, M., Herscovitch, P., Sunderland, T. & Molchan, S.E. (1997) Lateralization and behavioral correlation of changes in regional cerebral blood flow with classical conditioning of the human eyeblink response. *J Neurophysiol* **77**, 2153–2163.

Schreurs, B.G., Sanchez-Andres, J.V. & Alkon, D.L. (1991) Learning-specific differences in Purkinje-cell dendrites of lobule HVI (Lobulus simplex): intracellular recording in a rabbit cerebellar slice. *Brain Res* **548**, 18–22.

Stork, O. & Welzl, H. (1999) Memory formation and the regulation of gene expression. *Cell Mol Life Sci* **55**, 575–592.

Stubley-Weatherly, L., Harding, J.W. & Wright, J.W. (1996) Effects of discrete kainic acid-induced hippocampal lesions on spatial and contextual learning and memory in rats. *Brain Res* **716**, 29–38.

Tudor, M., Akbarian, S., Chen, R.Z. & Jaenisch, R. (2002) Transcriptional profiling of a mouse model for Rett syndrome reveals subtle transcriptional changes in the brain. *Proc Natl Acad Sci U S A* **99**, 15536–15541.

Yao, P.J., Zhu, M., Pyun, E.I., Brooks, A.I., Therianos, S., Meyers, V.E. & Coleman, P.D. (2003) Defects in expression of genes related to synaptic vesicle trafficking in frontal cortex of Alzheimer's disease. *Neurobiol Dis* **12**, 97–109.

Yeo, C.H., Hardiman, M.J. & Glickstein, M. (1985) Classical conditioning of the nictitating membrane response of the rabbit. II. Lesions of the cerebellar cortex. *Exp Brain Res* **60**, 99–113.

Acknowledgments

We gratefully acknowledge Alfia Corsino, Maria Patrizia D'Angelo and Francesco Marino for their administrative and technical support. This work was partly sponsored by grants of the Italian Ministry of Health and the Italian Ministry of Education University and research to SC.

1601183x, 2006, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/j.1601-183X.2005.00159.x, Wiley Online Library on [13/10/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

RESEARCH ARTICLE
Therapeutics and Prevention

Comprehensive Bioinformatic Assessments of the Variability of *Neisseria gonorrhoeae* Vaccine Candidates

Benjamin I. Baarda,[a] Ryszard A. Zielke,[a] Alaina K. Holm,[a] Aleksandra E. Sikora[a,b]

[a]Department of Pharmaceutical Sciences, College of Pharmacy, Oregon State University, Corvallis, Oregon, USA
[b]Vaccine and Gene Therapy Institute, Oregon Health and Science University, Beaverton, Oregon, USA

ABSTRACT A protective vaccine is the only viable way to stop the spread of gonorrhea in the face of rising antibiotic resistance. However, the notorious phase and antigenic variation of *Neisseria gonorrhoeae* surface proteins remains one of the challenges in vaccine development. To facilitate vaccine advancement efforts, we carried out comprehensive bioinformatic analyses of sequence variation by comparing 34 gonorrhea antigen candidates among $>$5,000 clinical *N. gonorrhoeae* isolates deposited in the *Neisseria* PubMLST database. Eight protein antigens showed exceptional conservation by having a single allele variant distributed in $>$80% of isolates. An additional 18 vaccine candidates were represented by $\leq$3 alleles in $>$50% of *N. gonorrhoeae* isolates globally. Phylogenetic analyses highlighted closely related antigen variants and additionally showed that AniA and FetB were the closest between *N. gonorrhoeae* and *N. meningitidis*. Up to 44% of *N. meningitidis* alleles for both antigens have premature stop codons, suggesting differential expression. Mapping polymorphisms to the available three-dimensional structures of 12 antigens revealed low-frequency surface polymorphisms. PorB and TbpB possessed numerous high-prevalence polymorphic sites. While TbpA was also highly variable, conserved loops were nonetheless identified. A high degree of sequence conservation, the distribution of a single antigen variant among *N. gonorrhoeae* strains globally, or low-frequency sequence polymorphisms in surface loops make ACP, AniA, BamA, BamE, MtrE, NspA, NGO0778, NGO1251, NGO1985, OpcA, PldA, Slam2, and ZnuD promising candidates for a gonorrhea vaccine. Finally, the commonly used *N. gonorrhoeae* FA1090 strain emerges as a vaccine prototype, as it carries antigen sequence types identical to the most broadly distributed antigen variants.

IMPORTANCE *Neisseria gonorrhoeae*, the Gram-negative bacterium responsible for the sexually transmitted infection gonorrhea, is categorized as a high-priority pathogen for research and development efforts. *N. gonorrhoeae*'s "superbug" status, its high morbidity, and the serious health impact associated with gonorrhea highlight the importance of vaccine development. One of the longstanding barriers to developing an effective vaccine against *N. gonorrhoeae* is the remarkable variability of surface-exposed antigens. In this report, we addressed this roadblock by applying extensive bioinformatic analyses to 34 gonorrhea antigen candidates among $>$5,000 clinical *N. gonorrhoeae* isolates. Our studies are important, as they reveal promising, conserved gonorrhea vaccine candidates and aid structural vaccinology. Moreover, these approaches are broadly applicable to other infectious diseases where surface antigen variability impedes successful vaccine design.

KEYWORDS gonorrhea, protein subunit vaccine, antigen variability, PubMLST, bioinformatics, crystal structure, phylogenetics, *Neisseria gonorrhoeae*, antigenic variation, phylogenetic analysis, subunit, vaccine

Citation Baarda BI, Zielke RA, Holm AK, Sikora AE. 2021. Comprehensive bioinformatic assessments of the variability of *Neisseria gonorrhoeae* vaccine candidates. mSphere 6:e00977-20. https://doi.org/10.1128/mSphere.00977-20.

Editor Drusilla L. Burns, U.S. Food and Drug Administration

Copyright © 2021 Baarda et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International license.

Address correspondence to Aleksandra E. Sikora, Aleksandra.Sikora@oregonstate.edu.

In-depth bioinformatics profiling of gonorrhea antigens' variability for rational vaccine development

Received 23 September 2020
Accepted 8 January 2021
Published 3 February 2021

Development of vaccines against many infectious diseases and cancers has been hampered by immune system evasion strategies, including variability of surface antigens (1–3). The pathogenic *Neisseria* species, *Neisseria gonorrhoeae* and *Neisseria meningitidis*, are notorious for their ability to alter their surface proteome. Both *N. gonorrhoeae* and *N. meningitidis* have three important surface structures capable of phase and antigenic variation (2), namely, lipooligosaccharide (LOS) glycan (4), opacity proteins (5), and type IV pilus (6), as well as 50 to 100 genes that undergo solely phase variation (2, 7). The success of vaccines against *N. meningitidis* serogroup B infections demonstrated that creative vaccine development strategies are necessary to circumvent the difficulties posed by antigen variability. Through reverse and rational vaccinology approaches, two meningococcal subunit vaccines, Bexsero (GSK; also known as 4CMenB) (8, 9) and Trumenba (Pfizer) (8, 10, 11), were developed and entered the market in 2014. Bioinformatic assessment of factor H binding protein conservation revealed its two immunologically distinct subfamilies, which prompted the inclusion of a representative of each subfamily in the final Trumenba vaccine (10–12). These successes suggest that a similar approach could benefit gonorrhea vaccine research, which has been focusing on a handful of antigens for years (13).

N. gonorrhoeae is classified as one of the five "urgent" antibiotic-resistant threats in the United States by the Centers for Disease Control and Prevention (CDC) (14). Until recently, the sole treatment regimen approved by the CDC was azithromycin combined with ceftriaxone (15, 16). However, this recommendation was changed to a single dose of ceftriaxone based on the increased occurrence of azithromycin resistance, the low incidence of ceftriaxone resistance, and antimicrobial stewardship concerns (17). A vaccine appears as the only sustainable line of defense against the detrimental health effects caused by the gonococcus, which include urethritis, cervicitis, proctitis, neonatal conjunctivitis and blindness, mucosal infection of the rectum and pharynx, and the facilitation of HIV transmission (18–21). The common hallmarks of urethral infection in men are purulent exudate and painful urination, while women may experience mild and nonspecific symptoms as well as dysuria and vaginal discharge (22, 23). However, many men and most women are asymptomatic (24, 25). If left untreated, gonorrhea can result in complications, including epididymitis, endometritis, pelvic inflammatory disease, and ectopic pregnancy (26, 27). In many countries, gonorrhea case rates are increasing and disparities in the overall disease rate and burden persist among sexual, gender, racial, and ethnic minorities as well as resource-limited populations (14, 28, 29). The CDC, its European counterpart, the ECDC, and the World Health Organization (WHO) have all emphasized the dire threat posed by drug-resistant gonorrhea (26, 27, 30). Vaccine development efforts must be expanded and accelerated to slow the spread of this disease (26).

Ever more powerful and useful genomics, proteomics, and bioinformatics tools are instrumental for broadening the repertoire of gonorrhea vaccine candidates and facilitating vaccine development. For instance, quantitative proteomic and immunoproteomic studies identified over 20 new gonorrhea vaccine candidates. These proteins are present at similar levels in *N. gonorrhoeae* cell envelopes of 19 diverse isolates (31, 32), including the 2016 WHO reference strains, and in native outer membrane vesicles (OMVs) of four commonly used laboratory strains (32). Further, high-throughput proteomics revealed their expression profiles under conditions relevant to host infection, including iron starvation, exposure to human serum, or anaerobiosis (33). These studies combined with data from omics profiling of *N. gonorrhoeae* biofilms and during host infection provide invaluable insights about expression of all gonorrhea antigens currently being pursued in preclinical studies (13, 34–39). However, information about their antigenic variability globally, which is critically needed to identify conserved antigens with potential to develop broadly protective immunity against the highly variable *N. gonorrhoeae*, remains scarce. To address this gap, in this report we assessed the conservation of 34 candidate vaccine antigens discovered through different traditional (13) and proteomics-based reverse vaccinology efforts (31–33). We examined sequence variation and phylogenetic relationships among alleles for a single protein using sequence data from >5,000 *N. gonorrhoeae* isolates in the publicly available *Neisseria* multilocus sequence typing database (*Neisseria* PubMLST; https://pubmlst.org/organisms/neisseria-spp, sited at the University of Oxford) (40). This database contains sequence data from over 60,000 isolates of pathogenic and commensal *Neisseria* species collected worldwide, providing a rich resource for antigen diversity mining. We additionally mapped amino acid polymorphic sites to available antigen crystal structures. This approach provides a visual representation of protein conservation and illuminates conserved surface loops that could be incorporated into a vaccine delivery platform, thereby facilitating structural vaccinology. Polymorphic amino acids were mapped based on their prevalence in the population, rather than on the raw number of sequences available, to give more relevant insights about allele distribution. Finally, we also performed comparative analyses between each gonococcal protein antigen and its meningococcal homolog, to determine the level of conservation between the two species and to identify candidates for inclusion in a cross-protective vaccine against both *N. gonorrhoeae* and *N. meningitidis*.

The approaches we describe for *N. gonorrhoeae* are broadly applicable, as the PubMLST family of databases is home to sequence data from over 100 species and genera, including prokaryotic and eukaryotic pathogens (40, 41). New isolates are added to the database frequently, providing an up-to-date understanding of global protein variation.

RESULTS AND DISCUSSION

Antigen selection and workflow for bioinformatics mining. We selected the 34 gonorrhea protein antigens that were discovered through traditional approaches (13) and proteome-based reverse vaccinology studies (31–33). Excluded were the highly polymorphic opacity-associated proteins, candidates identified experimentally to be periplasmic, MetQ and MlaA (reported in references 38 and 42), and the small lipoprotein NGO2054 (33, 43) that has inconsistent annotations in PubMLST (e.g., annotated independently, as part of NGO2055, or lacks annotation). For comprehensive bioinformatics mining of *N. gonorrhoeae* antigens, we have developed the workflow presented in Fig. 1. The antigen sequences were downloaded according to the detailed instructions we previously described (41). Eight of the 34 antigens, including LbpA, PilQ, IgA2 protease, CsgG, MtrE, OpcA, TbpB, and ZnuD, contained various numbers of alleles disrupted by premature termination codon (see Table S1 in the supplemental material). The highest numbers of these alleles, 31, 18, and 16, were identified for LbpA, PilQ, and IgA2 protease, respectively. The presence of a premature stop codon may suggest phase variation events and most commonly results in loss of protein expression, a nonfunctional truncated protein, or, rarely, a protein with altered function (44). Therefore, alleles containing premature stop codons (89 in total) were excluded from further analysis. The *N. gonorrhoeae* strain FA1090 was used as a reference to identify the single nucleotide polymorphisms (SNPs) and single amino acid polymorphisms (SAAPs) in the obtained antigen sequences (Data Set S1). For each antigen, we mapped SAAPs identified in the most frequent antigen variant to the corresponding allele in *N. gonorrhoeae* FA1090 (Table 1). Finally, based on the prevalence data available from PubMLST, the frequencies of each SAAP were calculated, and the most common polymorphisms (found in ≥1% of the global population) were mapped to available protein crystal structures, either from *N. gonorrhoeae* or the closely related *N. meningitidis* (see Fig. 6 and 7).

NGO0778, SliC, PldA, NGO1985, BamE, ACP, Slam2, and NGO1251 are highly conserved gonorrhea vaccine candidates. To illustrate the distribution of each antigen nucleotide sequence variants in the global *N. gonorrhoeae* population, we graphed the percentage of isolates associated with each antigen allele (Fig. 2). Remarkably, over 90% of isolates carry an identical allele of NGO0778, SliC, PldA, NGO1985, and BamE. Further, between 80% and 90% of *N. gonorrhoeae* strains have a sole nucleotide sequence for acyl carrier protein (ACP), Slam2, and NGO1251. For 18 additional

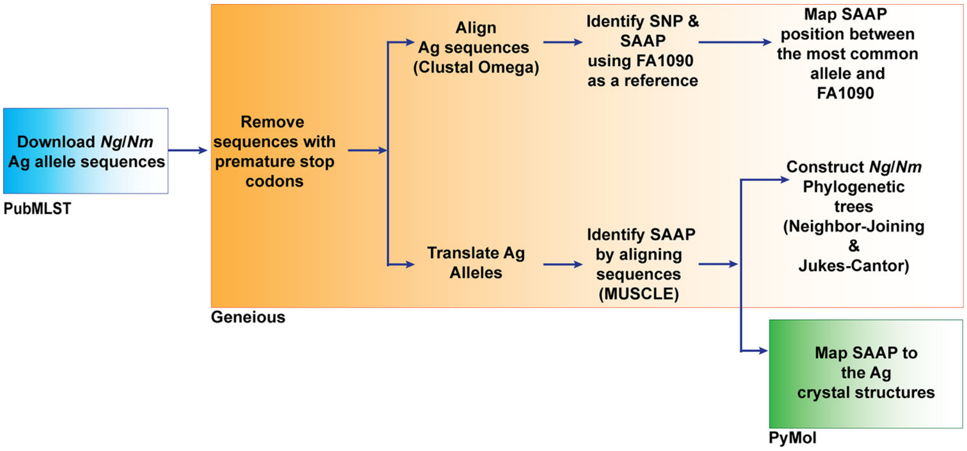

Download Ng/Nm Ag allele sequences
PubMLST
Remove sequences with premature stop codons
Align Ag sequences (Clustal Omega)
Identify SNP & SAAP using FA1090 as a reference
Map SAAP position between the most common allele and FA1090
Translate Ag Alleles
Identify SAAP by aligning sequences (MUSCLE)
Construct Ng/Nm Phylogenetic trees (Neighbor-Joining & Jukes-Cantor)
Geneious
Map SAAP to the Ag crystal structures
PyMol

FIG 1 Workflow for *N. gonorrhoeae* antigen variability mining. The *N. gonorrhoeae* and *N. meningitidis* (*Ng*/*Nm*) antigen (Ag) nucleotide sequences were downloaded from the PubMLST database (blue box). Geneious Prime 2020.1.2 software was subsequently used to perform analyses described below and outlined in the orange box. Sequences containing premature stop codons were eliminated. Nucleotide sequences were aligned using the Clustal Omega algorithm. To identify SNPs and SAAPs from the deduced amino acid sequence compared to the FA1090 allele sequence as a prototype, the Annotate & Predict tool was applied. In parallel, nucleotide sequences were translated to their respective amino acid sequences and aligned using the MUSCLE algorithm, followed by generation of neighbor-joining phylogenetic trees with the Jukes-Cantor genetic distance model. For proteins with available structural data, the proportion of isolates associated with each SAAP compared to the most common amino acid was calculated. Polymorphisms present in ≥1% of isolates were mapped to the protein three-dimensional structure using PyMOL (green box). SNP, single nucleotide polymorphism; SAAP, single amino acid polymorphism.

antigens—LprI, NGO1559, PilN, BamA, CsgG, NGO0425, Ng-MIP, ZnuD, LolB, NspA, OmpU, TamA, LptD, OpcA, PilQ, NGO1344, AniA, and MtrE—three or fewer alleles encompass >50% of *N. gonorrhoeae* isolates worldwide.

As expected, at the protein level, less unique SAAPs are present in the population for each antigen due to the presence of synonymous mutations. Thus, fewer protein variants would need to be included in a vaccine than the number of alleles would suggest (Table 1; Data Set S1). The most highly conserved antigen was BamE, with eight alleles and six distinctive amino acid sequences (Table 1; Data Set S1). Eight different amino acid sequences were present for SliC and NGO1251, whereas 10, 10, and 14 distinct amino acid sequence variants occurred for NGO1985, NGO0778, and ACP, respectively. The least conserved antigen, with 1,262 alleles and 1,229 unique amino acid sequences, was PorB. The dramatic difference in conservation between antigens illustrates the importance of this approach to evaluate the variability of candidate antigens before their testing in preclinical studies and inclusion in a vaccine. Our results showed that the FA1090 amino acid sequence is the most prevalent globally in 10 of the 34 antigens analyzed and diverges from the most common amino acid sequence by one or two polymorphisms in an additional 13 antigens (Table 1). The sole exception is LbpA, which has a premature stop codon in the FA1090 allele. Therefore, for this locus we used as the reference sequence the most common LbpA allele 16, which is carried, for instance, by the 2016 WHO reference strains WHO-M, WHO-P, WHO-U, and WHO-Y (45).

Cumulatively, these results suggest that the eight well-conserved proteins with >80% global homogeneity have the potential to protect against a wide range of *N. gonorrhoeae* strains if they are strong immunogens that are expressed on the bacterial cell surface during host infection. A combination of two or more conserved antigens could increase the vaccine coverage and effectiveness in comparison to a single-component vaccine. An example of this strategy is the *N. meningitidis* serogroup B vaccine Trumenba (Pfizer), which includes two representative factor H binding protein (fHbp) variants to provide broader coverage (10, 11). Finally, the antigen conservation data

TABLE 1 Summary of antigen polymorphism analyses[a]

Antigen (length range; mean mol wt)	No. of alleles	No. of SNPs	No. of unique amino acid sequences	No. of SAAPs	Position(s) of polymorphism(s) between FA1090 and most common allele (residue no. FA1090$_{aa}$ » allele$_{aa}$)[b]
ACP (123–124 aa; 13.3 kDa)	22	41	14	16	20 A»−; 25 N»D
AniA (386–402 aa; 40.9 kDa)	101	233	52	59	323 N»S
BamA (791–793 aa; 88 kDa)	171	130	78	65	553 R»K; 554 K»Q; 585 S»L
BamE (125 aa; 13.9 kDa)	8	38	6	13	NA (FA1090 allele is most common)
CsgG (NGO0834) (220–224 aa; 23.7 kDa)	24[c]	44	19	25	119 V»I
FetB (NGO2092) (287–337 aa; 34.1 kDa)	163	77	71	35	6 A»T; 107 D»E; 193 G»−; 194 K»−
IgA1 protease (1,533–1,595 aa; 173.6 kDa)	297	815	248	324	142 S»G; 467 S»A; 468 N»D; 469 Q»K; 491 P»S; 492 D»N; 497 N»D; 567 N»D; 609 Q»V; 610 T»A; 616 I»V; 617 L»F; 620 S»Y; 827 Q»K
IgA2 protease (1,461–1,476 aa; 160.4 kDa)	174[c]	227[c]	116	106[c]	1271 G»E
LprI (339 aa; 36.8 kDa)	35	37	25	21	NA (no amino acid polymorphic sites between most common allele and FA1090 allele)
LptD (797–805 aa; 87.5 kDa)	183[d]	500[d]	85	196[d]	51 S»F; 97 E»Q; 101 K»Q; 446 H»Q
MafA (NGO1067) (313–321 aa; 34.6 kDa)	89	228	30	122	283 Q»R
MtrE (467–483 aa; 50.5 kDa)	64[c]	148[c]	36	40[c]	47 I»V; 454 V»A
NGO1985 (203 aa; 21.7 kDa)	14	37	10	16	NA (no amino acid polymorphic sites between most common allele and FA1090 allele)
NGO0425 (208–209 aa; 23.149 kDa	24	47	17	25	NA (no amino acid polymorphic sites between most common allele and FA1090 allele)
NGO0778 (144 aa; 16.1 kDa)	13	12	10	9	NA (FA1090 allele is most common)
NGO1559 (223–225 aa; 23.5 kDa)	32	30	23	22	7 F»S
NGO1251 (191–192 aa; 21.8 kDa)	18	29	8	14	NA (no amino acid polymorphic sites between most common allele and FA1090 allele)
NGO1344 (700–742 aa; 77.2 kDa)	119	266	50	150	NA (no amino acid polymorphic sites between most common allele and FA1090 allele)
NgMIP (NGO1225) (269–278 aa; 28.9 kDa)	36	51	21	21	49 S»G; 160 E»K
NspA (NGO0233) (173–175 aa; 18.4 kDa)	29[c,d]	22[c,d]	20	14	153 V»I
LbpA (881–952 aa; 105.7 kDa)	122	175	62	83	Internal stop codon for FA1090; SNP and nsSNP analysis done against most common allele
LolB (NGO0439) (193 aa; 21.2 kDa)	25	33	20	21	74 S»G
OmpU (NGO1688) (485–491 aa; 56.5 kDa)	165	200	123	119	37 C»R; 213 V»I; 270 V»A
OpcA (NGO0868) (262–264 aa; 28.9 kDa)	54[c]	27	30	18	100 E»K,116 I»L, 192 V»A
PilN (NGO0097) (199 aa; 22.2 kDa)	38	29	23	18	92 E»G
PldA (373–376 aa; 42 kDa)	27	37	16	16	NA (FA1090 allele is most common)
PorB (326–357 aa; 37.1 kDa)	1,262	507	1,229	269	18 M»T; 40 T»R; 41 D»E; 45 S»I; 46 K»G; 48 E»G; 54 A»S; 75 V»I; 89 T»S; 98 V»I; 120 G»K; 121 A»N; 187 Q»R; 189 S»N; 208 K»−; 209 I»−; 210 E»−; 211 Y»M; 212 D»E; 214 Q»Y; 215 T»A; 217 S»N; 242 V»A; 272 A»V; 295 S»D;
Slam2 (505–511 aa; 55.6 kDa)	53	483	41	181	NA (no amino acid polymorphic sites between most common allele and FA1090 allele)
SliC (122–126 aa; 13.5 kDa)	12	22	8	11	NA (FA1090 allele is most common)
TamA (610–616 aa; 67.7 kDa)	140	174	72	69	361 R»Q
TbpA (905–936 aa; 102.1 kDa)	638[c]	592[c]	573	217[c]	110 A»S; 247 V»A; 257 Y»H; 262 E»A; 265 K»E; 267 E»G; 268 G»S; 271 K»T; 274 A»T; 275 −»R; 276 −»P; 281 D»A; 356 E»A; 358 K»Q; 359 −»K; 360 K»Q; 361 Y»A; 364 I»L; 367 Y»N; 373 G»N; 374 R»H; 377 S»G; 380 I»F; 382 N»S; 386 G»N; 388 E»P; 421 T»A; 463 S»F; 509 D»N; 519 H»S; 530 P»A; 539 −»P; 540 −»D; 555 G»R; 561 G»R; 630 T»A; 665 G»D; 699 N»D; 712 A»V; 720 Q»G; 836 D»N; 887 A»G;

(Continued on next page)

TABLE 1 (Continued)

Antigen (length range; mean mol wt)	No. of alleles	No. of SNPs	No. of unique amino acid sequences	No. of SAAPs	Position(s) of polymorphism(s) between FA1090 and most common allele (residue no. $FA1090_{aa}$ » $allele_{aa}$)[b]
TbpB (691–703 aa; 74.7 kDa)	320[c]	1792[c]	311	219[c]	40 P»A; 74 P»R; 75 P»M; 76 S»A; 78 P»—; 79 K»—; 83 I»V; 84 R»K; 86 S»N; 88 G»S; 94 G»D; 100 N»E; 106 N»S; 112 S»T; 115 G»D; 116 E»G; 117 A»E; 118 P»T; 125 Q»—; 126 G»Q; 134 D»G; 149 Q»H; 151 G»R; 152 N»T; 153 T»K; 154 I»P; 155 K»E; 156 K»—; 157 D»—; 158 D»—; 159 S»T; 160 S»I; 161 S»D; 162 K»G; 163 I»K; 164 I»V; 165 E»T; 166 A»V; 168 N»S; 190 S»K; 192 E»M; 204 Q»K; 214 S»N; 216 F»S; 217 T»A; 219 S»P; 235 R»K; 238 E»D; 256 N»D; 258 E»G; 300 M»I; 301 V»A; 306 E»K; 307 N»A; 308 S»N; 309 K»E; 310 S»T; 312 Q»E; 330 Q»K; 341 N»D; 343 N»G; 373 N»G; 375 A»T; 379 S»P; 381 G»E; 382 N»—; 383 S»N; 384 N»K; 400 E»K; 424 S»P; 436 K»E; 439 G»K; 441 D»A; 443 T»I; 445 T»E; 449 M»T; 458 K»Q; 459 A»T; 460 Q»G; 461 T»M; 462 G»A; 463 A»T; 464 G»N; 466 M»V; 469 A»V; 471 D»N; 472 A»T; 475 V»G; 476 N»T; 477 G»S; 479 Q»K; 480 A»T; 481 G»—; 482 T»K; 483 K»T; 484 T»H; 488 E»Q; 505 E»K; 508 N»E; 510 V»A; 513 T»A; 514 V»G; 515 R»E; 516 N»S; 519 Q»R; 520 A»T; 522 A»V; 523 R»Q; 546 E»D; 547 Q»G; 548 K»N; 561 A»I; 568 G»R; 569 N»E; 572 D»N; 574 Q»E; 575 S»N; 611 D»E; 612 D»G; 623 N»D; 624 D»G; 632 S»N; 637 Y»H; 643 N»E; 645 E»K; 670 T»A; 671 K»E; 673 A»—; 674 Q»—; 675 E»—; 676 N»—; 679 A»T; 688 V»G; 700 Q»E
ZnuD, TdfJ (NGO1205) (763–764 aa; 85.7 kDa)	161[c]	93[c]	52	47	658 A»D

[a] Antigens are listed in alphabetical order. aa, amino acids.

[b] Residue number includes gaps only if present in one of the two sequences. Dashes indicate a single amino acid deletion (gap). NA, not applicable; nsSNP, nonsynonymous single-nucleotide polymorphisms.

[c] Includes only those alleles without an annotated premature stop codon.

[d] Excludes allele annotated as atypical.

further validate the choice of FA1090 as a vaccine antigen prototype and a challenge strain in immunization/challenge experiments (32, 33, 36, 38, 39, 41, 46).

Phylogenetic analyses of gonorrhea vaccine candidates reveal evolutionarily distinct groups. We constructed phylogenetic trees to determine the evolutionary relationship between alleles and to assess distinct *N. gonorrhoeae* antigen families (Fig. 3 and 4; Fig. S1 to S3). The results of our phylogenetic analyses reflect and expand upon our prior observations (Table 1 and Fig. 2) by illustrating the number of alleles and how closely related they are on the amino acid level. The majority of alleles for each antigen were closely related, with notable exceptions. Although it is the most highly conserved of the antigens examined, BamE has a single allele that is divergent from the rest and forms an outgroup (Fig. 3A). Similar single-allele outgroups are observed for CsgG (Fig. 3B), LptD (Fig. 3D), Ng-MIP (Fig. 3F), NGO0425 (Fig. 4A), NGO1251 (Fig. 4B), OmpU (Fig. 4C), PilQ (Fig. 4E), and TbpB (Fig. 4G). These outgroups represent a minimal proportion of *N. gonorrhoeae* isolates for each antigen. NGO0425 has the outgroup that encompasses the most isolates, at 13 (0.2% of the isolates with data for this locus).

Our results are also consistent with the separation of several antigens into multiple subfamilies. We observed distinct subfamilies for IgA1 protease (Fig. 3C), MafA (Fig. 3E), PorB (Fig. 4D), and TbpB (Fig. 4G), discussed below. In our analysis, IgA1 protease clusters into two evolutionarily distinct groups, which correlate with the established

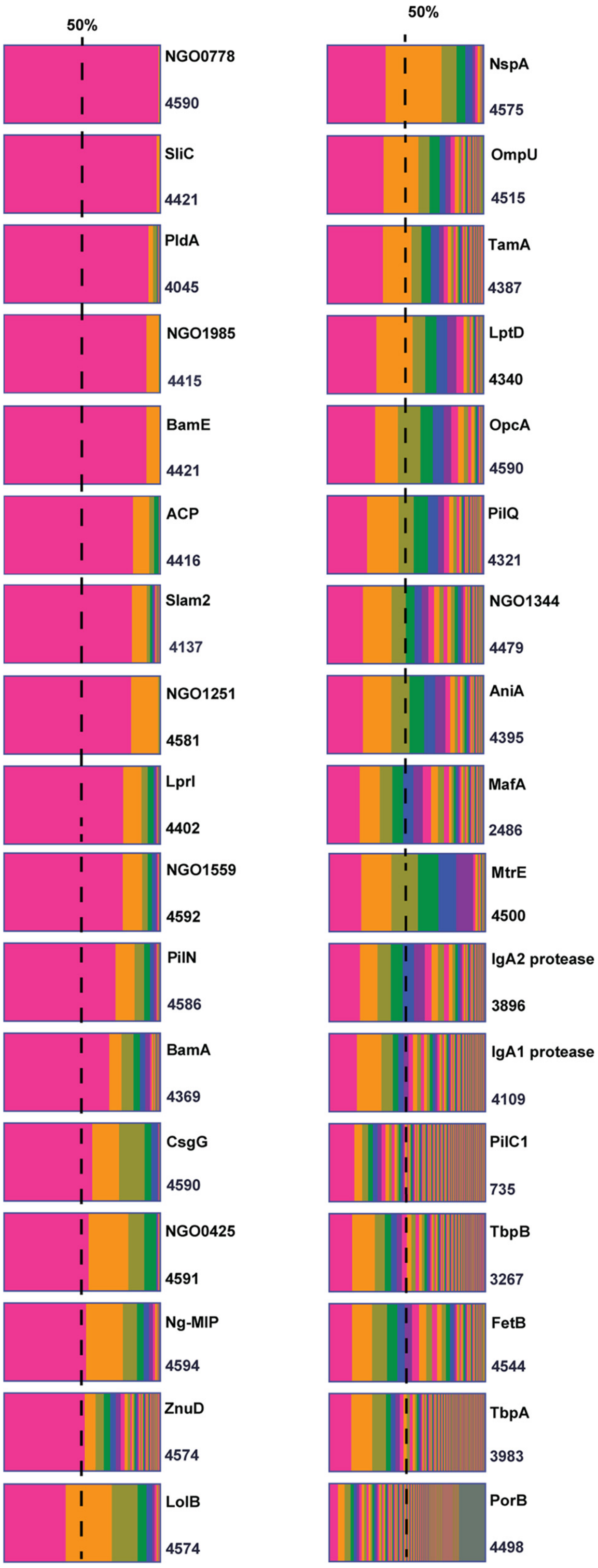


FIG 2 Distribution of each antigen nucleotide sequence variant in the global *N. gonorrhoeae* population. The percentages of *N. gonorrhoeae* isolates associated with each antigen allele were generated using GraphPad Prism 9. The total number of isolates with data for the antigen is indicated at the bottom right of each graph. The dotted line represents 50% of *N. gonorrhoeae* isolates. Antigens are arranged based on the prevalence of the most common allele in the *N. gonorrhoeae* global population.

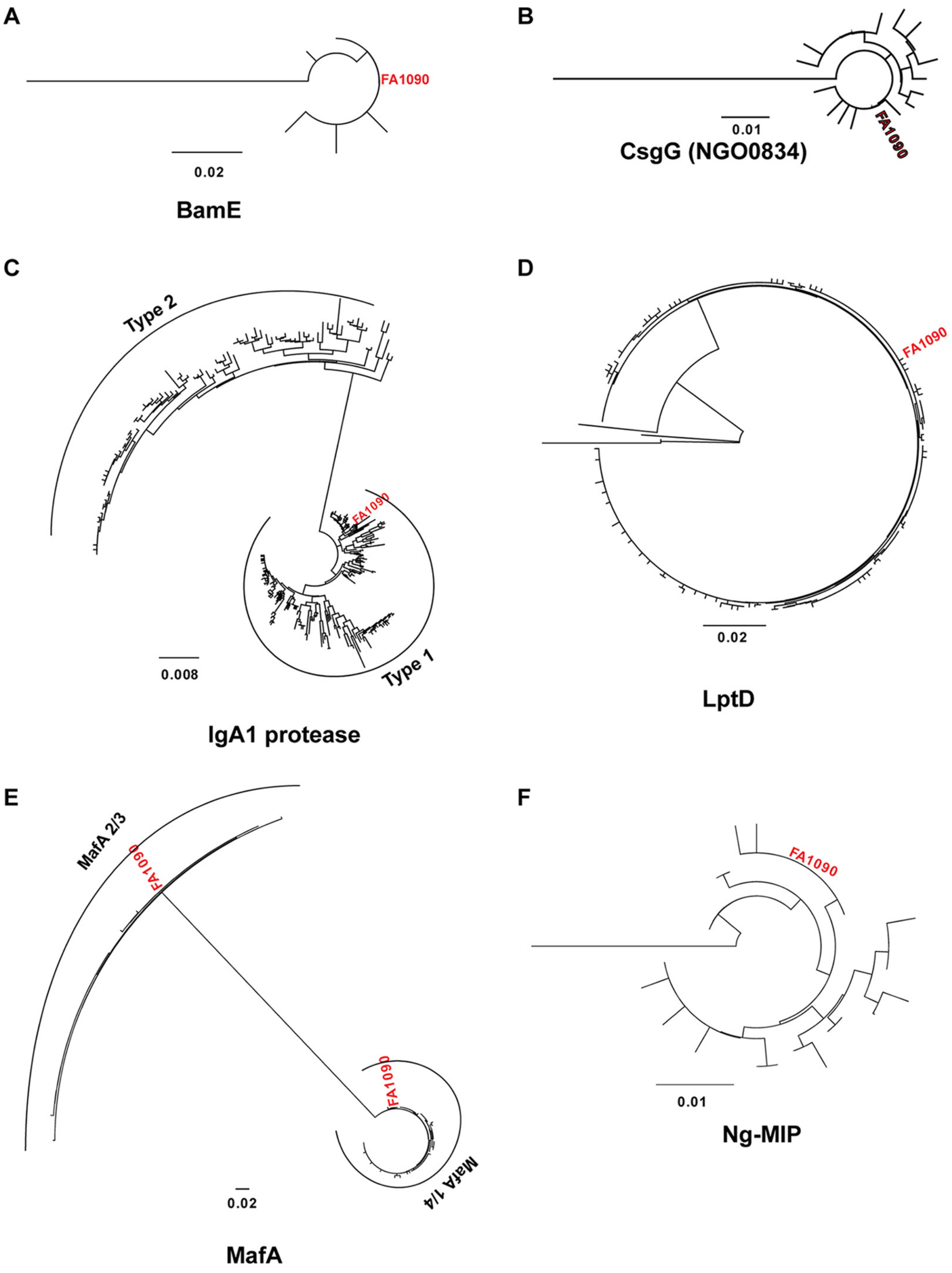

A
FA1090
0.02
BamE
B
0.01
CsgG (NGO0834)
FA1090
C
Type 2
FA1090
0.008
Type 1
IgA1 protease
D
FA1090
0.02
LptD
E
MafA 2/3
FA1090
FA1090
0.02
MafA 1/4
MafA
F
FA1090
0.01
Ng-MIP

FIG 3 Phylogenetic relationships of BamE, CsgG, IgA1 protease, LptD, MafA, and Ng-MIP among *N. gonorrhoeae* isolates. Neighbor-joining phylogenetic trees of antigens (as indicated below each tree) from *N. gonorrhoeae* were constructed in Geneious using the Jukes-Cantor genetic distance model. The FA1090 allele is designated in red text for each antigen. Subfamilies, when established, are indicated with curved lines. Antigens are organized in alphabetical order: BamE (A), CsgG (B), IgA1 protease (C), LptD (D), MafA (E), Ng-MIP (F).

classification into type 1 and type 2 IgA1 proteases (Fig. 3C) (47). Although both subfamilies function as autotransporter serine proteases that cleave the hinge region of the IgA1 heavy chain (48–52), they recognize different cleavage sites. Type 1 cleaves a proline-serine bond, and type 2 cleaves a proline-threonine bond immediately upstream of the type 1 recognition site (47). Further, the two variants have different secondary substrates (53–55), which may contribute to gonococcal virulence in ways that are not

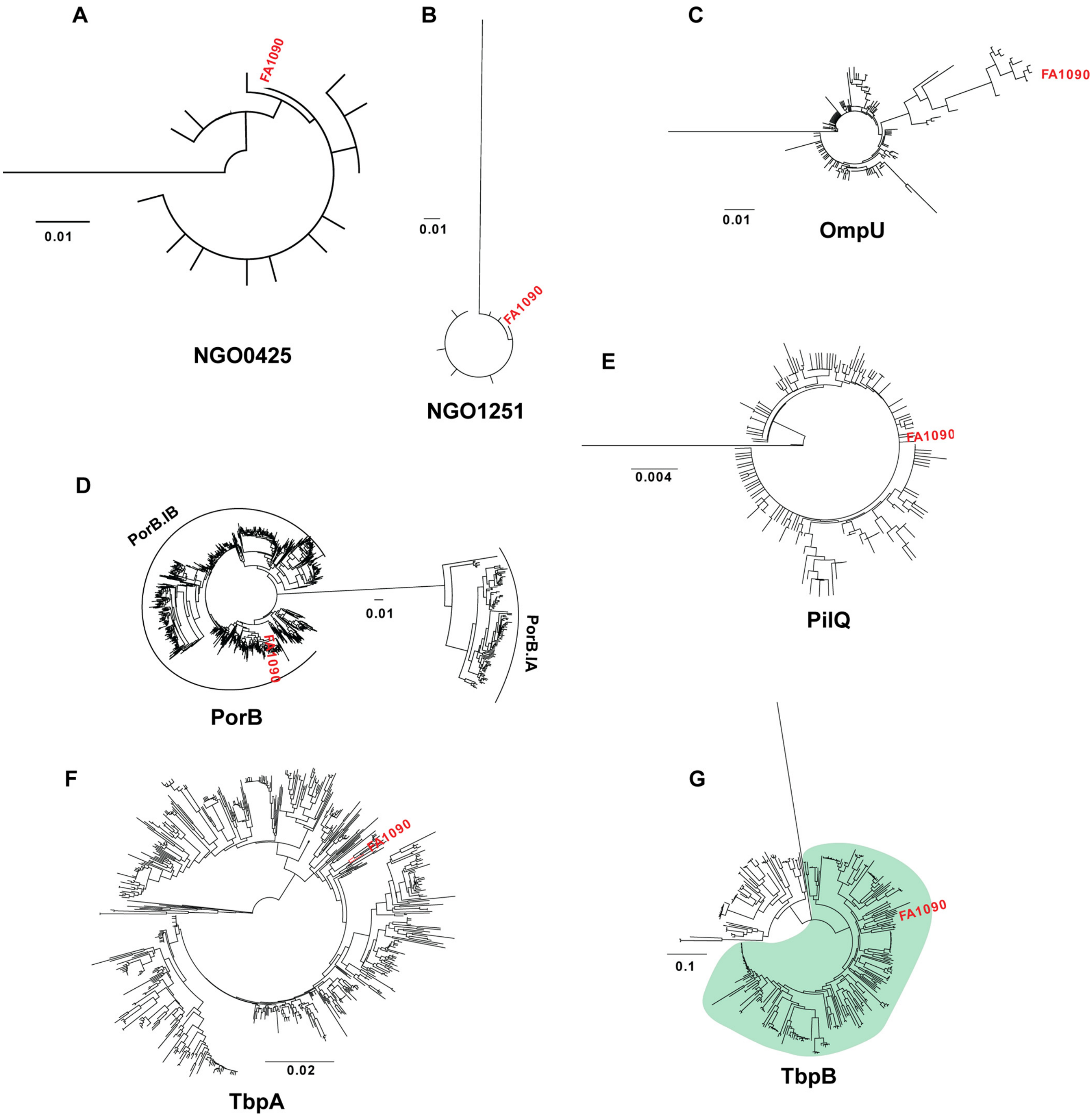

A
FA1090
0.01
NGO0425
B
0.01
FA1090
NGO1251
C
FA1090
0.01
OmpU
D
PorB.IB
0.01
FA1090
PorB.IA
PorB
E
0.004
FA1090
PilQ
F
FA1090
0.02
TbpA
G
0.1
FA1090
TbpB

FIG 4 Phylogenetic relationships of NGO425, NGO1251, OmpU, PorB, PilQ, TbpA, and TbpB among *N. gonorrhoeae* isolates. Phylogenetic trees of antigens were constructed in Geneious. The FA1090 allele for each antigen is designated in red text. Subfamilies, when established, are indicated with curved lines. The green-shaded region of TbpB phylogeny indicates the TbpB$_E$ subfamily. Antigens are listed in alphabetical order: NGO0425 (A), NGO1251 (B), OmpU (C), PorB (D), PilQ (E), TbpA (F), TbpB (G).

yet clear. Due to the evolutionary distance between the two protease types, a representative of each type will likely need to be considered for a vaccine, as *N. gonorrhoeae* isolates express one or the other but not both (47, 56).

MafA is an adhesin in the multiple adhesin family (*maf*) which binds to glycolipid cell receptors and mediates cell attachment (57). Up to five *maf* variants are encoded on horizontally acquired genomic islands in *N. gonorrhoeae* genomes (58). FA1090 possesses four copies of MafA: two pairs of identical loci, designated MafA 1/4 (NGO1067 and NGO1972) and MafA 2/3 (NGO1393 and NGO1584). Both pairs are returned as a single locus when the PubMLST database is queried. Our results show that the two MafA subtypes are evolutionarily distinct, while the alleles associated with each subtype are closely related (Fig. 3E). Therefore, a vaccine formulated with MafA should incorporate both subfamilies to account for isolates that may have one or the other.

The porin PorB (also known as protein I) is one of the most abundant proteins in the gonococcal cell envelope and acts primarily as a voltage-gated channel to facilitate ion exchange with the environment (59). *N. gonorrhoeae* isolates possess one copy of PorB in one of two allelic forms, PorB.IA or PorB.IB, which share only ~70% nucleotide sequence identity (60). The two protein forms are associated with different disease states: PorB.IA isolates are generally associated with disseminated infections, while PorB.IB-expressing strains usually cause localized urogenital infections (61, 62). PorB subtypes also contribute to fitness in the host in different ways (63–66). Consistent with the presence of two PorB subtypes in the *N. gonorrhoeae* population, this phylogenetic analysis split PorB alleles into two distinct clusters that correlate with PorB.IA (Fig. 4D, small cluster) and PorB.IB (Fig. 4D, large cluster). Although PorB.IA accounts for <10% of PorB sequences worldwide (426 of 4,498 isolates with PorB sequence data), representatives from both groups could be included in a vaccine, due to the difference in secondary function and extensive sequence diversity between the subtypes, as well as the dangerous disease phenotype associated with PorB.IA.

The outer membrane transferrin binding protein TbpA has previously been classified into two groups, based on the presence (D) or absence (N) of a deletion in variable region 3 (VR3) (67). However, our data do not support this classification. While we observe three clusters (two major and one minor) (Fig. 4F), both major groups have alleles with and without deletions in VR3. Instead, sequence differences in VR2, and to a lesser extent in VR1, appear to be the primary drivers of differentiation between the two main TbpA clusters. The possible reason for the differences observed between these reports is the small number of sequences analyzed in the previous study. Pajón et al. had access to only nine TbpA sequences in 1997 (67), while we compared TbpA variation among 638 allele sequences derived from nearly 4,000 isolates.

Finally, lipoprotein TbpB, which is involved in scavenging iron from human transferrin in conjunction with TbpA (68, 69), can be broadly classified into two isotypes based on protein size. Both *N. gonorrhoeae* and hyperinvasive meningococcal strains have type II TbpB (70). However, sequence analyses with 48 *Neisseria* isolates, including both pathogenic and commensal species, indicated that the two TbpB isotypes could be further divided into five families, from A to E. *N. gonorrhoeae* isolates are found in the TbpB$_C$, along with *N. meningitidis* and commensal species. TbpB$_E$ is composed exclusively of *N. gonorrhoeae* TbpB (71). Consistent with these results, phenotypic analysis clustered *N. gonorrhoeae* TbpB alleles into two groups (Fig. 4G). The two clusters are not as evolutionarily divergent as other groups observed in this study, likely because they are both members of the type II family. TbpB$_E$ (Fig. 4G, green-shaded area) accounts for the majority of alleles (247/320 alleles) and 80% of isolates.

These phylogenetic analyses illustrate the extent of sequence variation among the global *N. gonorrhoeae* population and highlight antigens' multiple variants, which may need to be considered in a vaccine design to account for distantly related protein families.

AniA and FetB are multi-*Neisseria* vaccine candidates based on phylogenetic analyses. We subsequently used phylogenetic comparisons to assess whether any gonorrhea vaccine candidate antigens could potentially be considered in a multi-*Neisseria* vaccine to protect against gonorrhea and meningococcal sepsis and meningitis simultaneously. While NGO1559 was highly conserved in *N. gonorrhoeae* (Fig. 2), it had no *N. meningitidis* homolog. With rare exceptions, *N. gonorrhoeae* protein variants clustered separately from their meningococcal homologs (Fig. 5; Fig. S4 to S8). Occasional instances of possible horizontal gene transfer between the species, where a gonococcal variant bundled with an otherwise exclusively *N. meningitidis* group or vice versa, could be observed for several antigens, including LptD, OmpU, Slam2, TamA, and TbpA (Fig. S4 to S8).

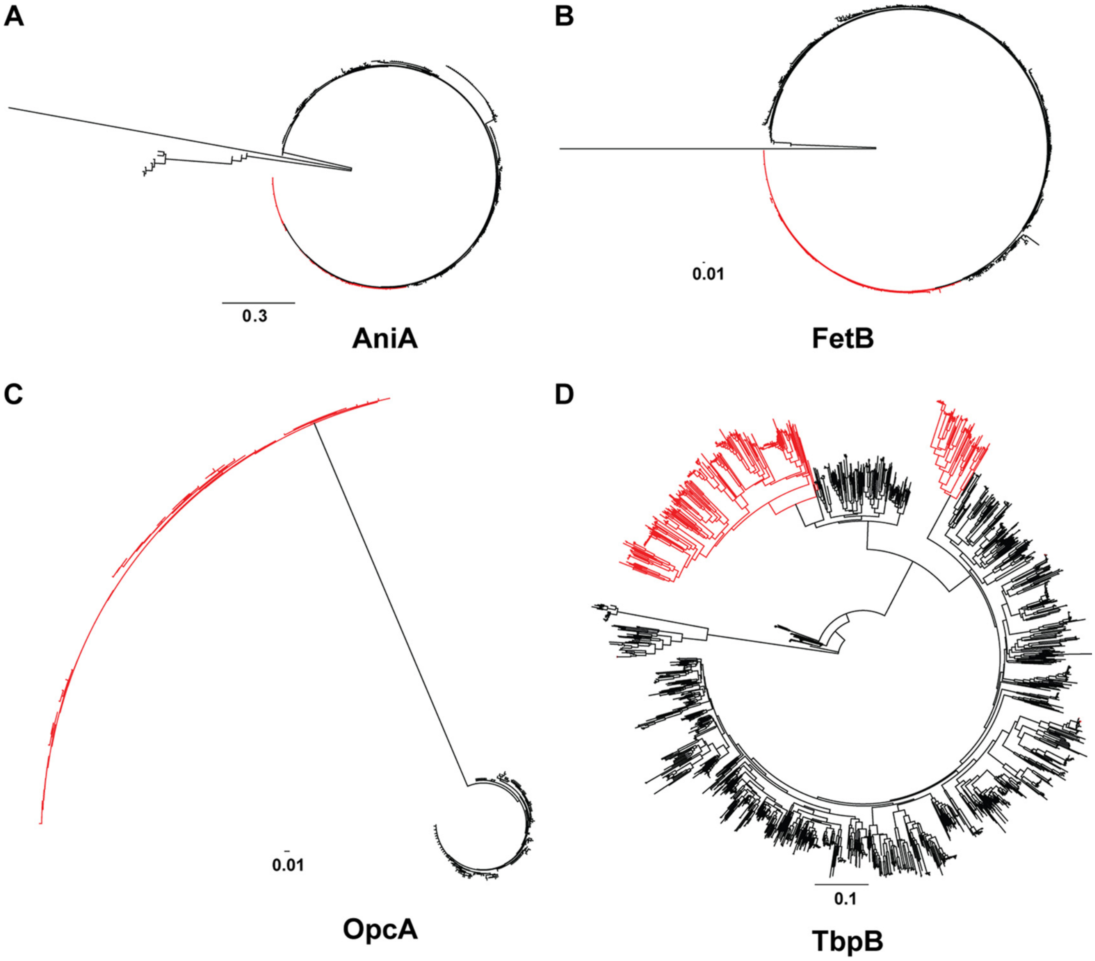


FIG 5 Phylogenetic relationships of AniaA, FetB, OpcA, and TbpB between *N. gonorrhoeae* and *N. meningitidis*. Neighbor-joining phylogenetic trees of indicated antigens from *N. gonorrhoeae* and *N. meningitidis* were constructed in Geneious using the Jukes-Cantor genetic distance model. *N. gonorrhoeae* alleles are designated using red branches. (A) AniaA; (B) FetB; (C) OpcA; (D) TbpB.

Based on phylogenetic comparisons between *N. gonorrhoeae* and *N. meningitidis*, the nitrite reductase AniA (Fig. 5A) and the component of the ferric enterobactin transport system FetB (Fig. 5B) seem the most useful candidates for inclusion in a multi-*Neisseria* vaccine. For both proteins, *N. gonorrhoeae* variants are in the same cluster as the *N. meningitidis* homologs or are in a closely related grouping. One consideration, however, is that 34% of *N. meningitidis* isolates have a frameshift mutation that abolishes AniA expression (72). Our analyses showed the existence of a higher percentage (43.8%) of AniA with premature stop codons in meningococcal isolates globally. Similarly, 34 alleles associated with 4,147 *N. meningitidis* isolates (28.8% of the population) encoded FetB sequences with premature stop codons.

The antigen subfamily separations seen within *N. gonorrhoeae* sequences (as discussed in a previous section) were further accentuated between *N. gonorrhoeae* and *N. meningitidis* isolates (Fig. 5C and D; Fig. S4 to S8). It was especially apparent for OpcA and TbpB. The gonococcal OpcA cluster was evolutionarily distinct from the *N. meningitidis* group (Fig. 5C). OpcA, an adhesin, was proposed as part of a DNA island imported into *N. gonorrhoeae* and *N. meningitidis* genomes from different species, based on the low homology between *N. gonorrhoeae* and *N. meningitidis* and the presence of a DNA uptake sequence upstream of the OpcA locus (73). From this standpoint, it appears as the least suitable candidate for a multi-*Neisseria* vaccine. For TbpB, *N.*

gonorrhoeae sequences formed two distinct groups completely separated from each other (Fig. 5D). The larger *N. gonorrhoeae* cluster corresponded to $\mathrm{TbpB_E}$ and did not have any *N. meningitidis* sequences associated with it (Fig. 5D). We noted that the majority of *N. gonorrhoeae* variants, although clustered separately from the *N. meningitidis* homologs, were moderately closely related to their meningococcal counterparts.

Together, these investigations suggest that while both AniA and FetB may be attractive candidates for a multi-*Neisseria* vaccine, additional, potentially species-specific antigens will need to be included to compensate for the lack of these proteins in substantial proportions of the *N. meningitidis* population.

Polymorphism mapping reveals conserved regions in vaccine antigens. To map antigen polymorphic sites, we used the available eight crystal structures of β-barrel-outer membrane proteins (OMPs) (Fig. 6) and four lipoproteins (Fig. 7) from either *N. gonorrhoeae* or the closely related *N. meningitidis*. In the past, we have performed similar analyses for the trimeric outer membrane channel component of the multiple transferable resistance (*mtr*) and fatty acid resistance (*far*) efflux pump systems, MtrE (74, 75), and the central component of the β-barrel assembly machinery (BAM) complex, BamA (41). Herein, however, we calculated the prevalence of each polymorphism within the *N. gonorrhoeae* population. This strategy allows the visual demonstration of the polymorphism frequency, in addition to its location, to determine whether several protein variants should be considered for a broad vaccine coverage (Fig. 6 and 7).

Importantly, only two low-prevalence polymorphisms (1 to 10%) were observed in surface loops of BamA (Fig. 6A). The highest-prevalence polymorphism, found in 20 to 30% of the *N. gonorrhoeae* population, was present in one of the periplasmic polypeptide transport-associated (POTRA) domains. An additional seven low-prevalence polymorphisms were found throughout the BamA β-barrel and POTRA domains. Remarkably, over 99% of *N. gonorrhoeae* isolates did not have any surface-exposed polymorphisms in MtrE (Fig. 6B). Per subunit, five polymorphisms (four in 10 to 20% of the population and one in 30 to 40%) are found in the equatorial domain of the periplasmic α-barrel, while a single high-prevalence polymorphism (40 to 50%) is present at the base of the periplasmic channel facing the pore's interior (Fig. 6B). Similarly, neisserial surface protein A (NspA), which contributes to *N. gonorrhoeae* serum resistance by binding factor H and factor H-like protein 1 (76), has a single low-prevalence polymorphism (1 to 10% of isolates) in one of its surface loops (Fig. 6C). Three more polymorphisms are present in the β-barrel portion of the protein; two are in 1 to 10% of isolates, and the third is in 30 to 40% (Fig. 6C). The adhesin OpcA exhibits three polymorphisms that are present in 10 to 20% of the population, two in a surface loop and one in the β-barrel portion, and one surface loop polymorphism that is present in 30 to 40% of the population (Fig. 6D). The dodecameric OMP antigen PilQ (77, 78) exhibited a single polymorphic site per subunit, within the structural data available for the periplasmic domains of *N. meningitidis* PilQ (78), that diverged in 20 to 30% of isolates (Fig. 6E). Further, no three-dimensional data are available for PorB.IB, despite its being the more prevalent subtype. However, mapping PorB.IA variants to the structure revealed extensive differences in all surface-exposed loops and, to a smaller degree, within the β-barrel itself (Fig. 6F). TbpA was highly variable, especially in a series of extracellular loops that extend ~60 Å above the cell surface (Fig. 6G) (79). This portion included eight sites that diverged from the most common sequence in 50 to 70% of isolates, along with five of the six sites that are polymorphic in 40 to 50% of the population. However, with the exception of D722 (polymorphic in 1 to 10% of isolates), the residues responsible for binding human transferrin were conserved in >99% of the *N. gonorrhoeae* population (80). In contrast, the zinc uptake protein ZnuD (TdfJ) was highly conserved and had only two surface loop polymorphisms found in 1 to 10% of the population within regions of the protein with available structural data (Fig. 6H).

Finally, ACP showed remarkable conservation of the four lipoprotein vaccine candidates with existing structural information (Fig. 7). The three-dimensional structure of *N. gonorrhoeae* ACP (81) did not incorporate any polymorphisms present in ≥1% of the

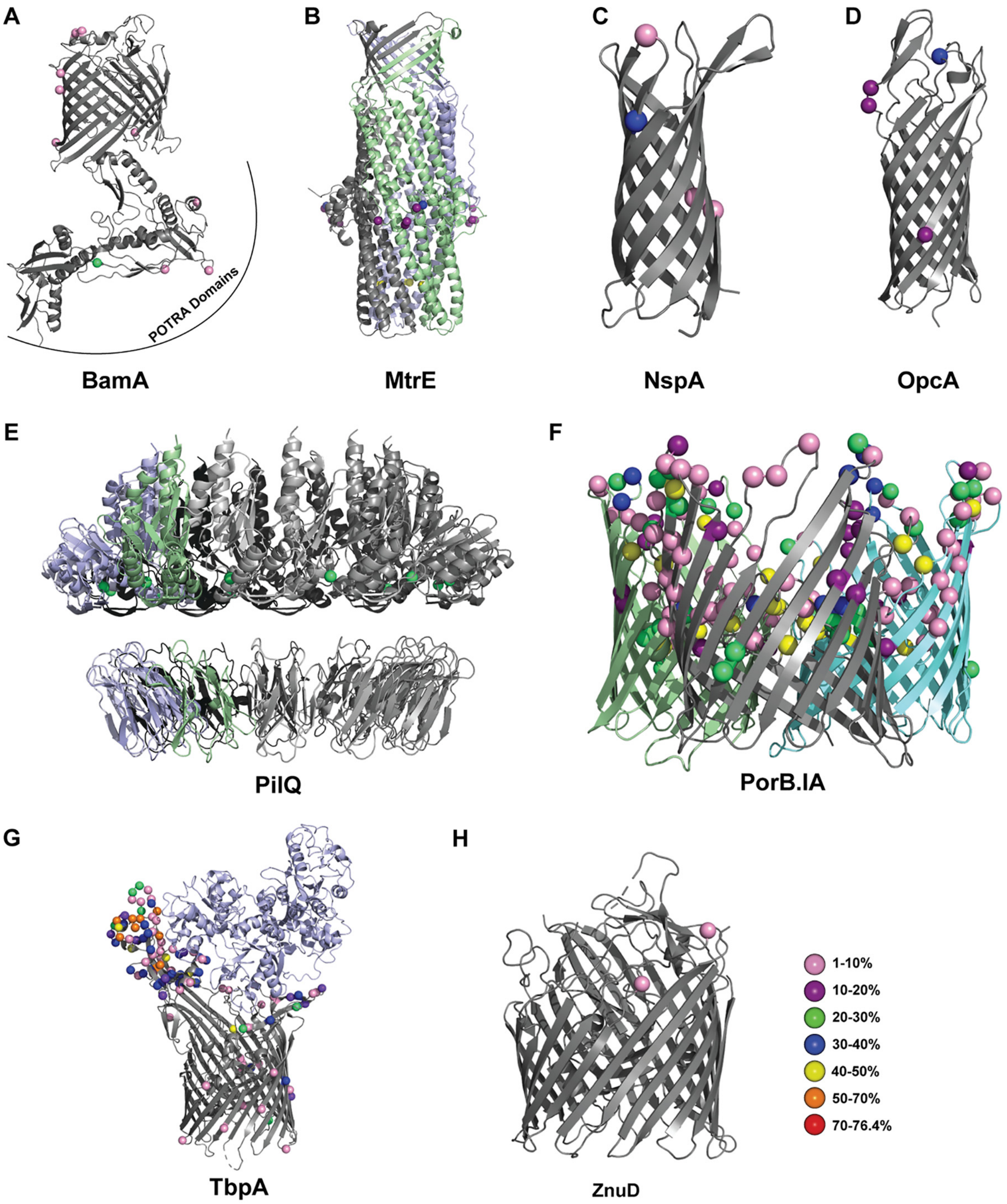


FIG 6 Polymorphism mapping to β-barrel outer membrane protein vaccine candidates. Polymorphic amino acid sites, weighted by the proportion of isolates associated with each polymorphism, were mapped to the structures of BamA (PDB accession no. 4K3B) (A), MtrE (4MT0) (B), NspA (1P4T) (C), OpcA (1K24) (D), PilQ (4AV2) (E), PorB.IA (4AUI) (F), TbpA (3V8X) (G), and ZnuD (4RVW) (H). When applicable, the biologically relevant multimers are shown, with individual monomers in different colors. Human transferrin associated with TbpA is shown in light blue. The polymorphic site color coding used in the figure is as follows: pink, 1 to 10%; purple, 10 to 20%; green, 20 to 30%; blue, 30 to 40%; yellow, 40 to 50%; orange, 50 to 70%; red, 70 to 76.4%.

population (Fig. 7A). Each subunit of the AniA trimer had four low-prevalence polymorphic sites (1 to 10%) and two sites divergent in 10 to 20% of the population (Fig. 7B). Each subunit of the BamE dimer possessed a single low-prevalence site (1 to 10% of the population) (Fig. 7C). TbpB was the most highly polymorphic lipoprotein examined,

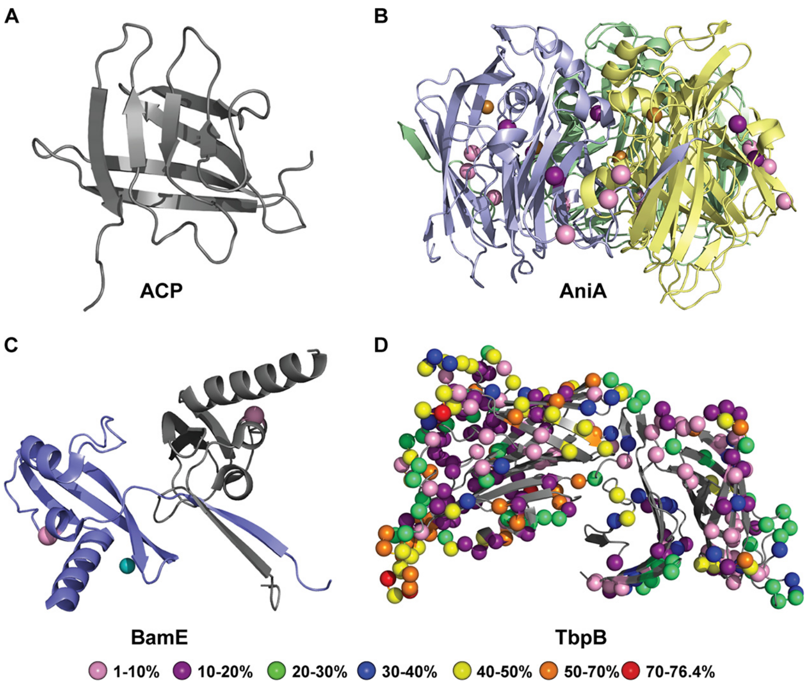


FIG 7 Polymorphism mapping to lipoprotein vaccine candidates. Polymorphic sites, weighted by the proportion of isolates associated with each polymorphism, were mapped to the structures of ACP (6GQ4) (A), AniA (5TB7) (B), BamE (5WAM) (C), and TbpB (3V8U) (D). When applicable, the biologically relevant multimers are shown, with individual monomers in different colors. Copper and zinc ions associated with AniA and BamE are designated by copper-colored and cyan spheres, respectively. Polymorphic site color coding is as follows: pink, 1 to 10%; purple, 10 to 20%; green, 20 to 30%; blue, 30 to 40%; yellow, 40 to 50%; orange, 50 to 70%; red, 70 to 76.4%.

with 379 polymorphic sites present in ≥1% of the global *N. gonorrhoeae* population (Fig. 7D). One hypervariable site diverged from the most common sequence in 76.4% of isolates. Including this site, four positions were polymorphic in >70% of the population, and 31 variants were present in 50 to 70% of isolates.

Cumulatively, these analyses showed that with the exception of PorB, TbpA, and TbpB, all of the OMPs and lipoproteins examined showed relatively few polymorphisms, particularly in the surface-exposed loops that are the most important from the vaccine design standpoint. Thus, BamA, MtrE, NspA, OpcA, ZnuD, ACP, AniA, and BamE appear as promising gonorrhea antigens. Several loops within the TbpA pore are highly conserved and could represent attractive targets for a vaccine that employed recombinant protein loops rather than the entire protein. A vaccine comprising the extracellular TbpA epitopes that elicits functional antibody responses targeting the transferrin binding pocket and occluding the pore might inhibit TbpA function. Similarly, thorough consideration and design would be necessary for TbpB. Indeed, a structure-based design to develop chimeric antigens to circumvent pathogen diversity has been successful in experimental vaccines against both bacterial and viral pathogens (82–85).

Conclusions. This is the first large-scale assessment of gonorrhea vaccine antigen variability. Evaluation of allele distributions revealed that >50% of *N. gonorrhoeae* isolates worldwide had three or fewer alleles for 23 antigens (Fig. 2). Among those, eight were exceptionally well conserved, with a single allele accounting for >80% of global isolates. Generation of phylogenetic trees indicated that the majority of antigens had closely related alleles and also revealed distinct subfamilies for IgA1 protease, MafA, PorB, and TbpB, consistent with classifications established in the literature (Fig. 3). Importantly, mapping polymorphic sites to structural data for 12 antigens showed that the majority of surface-exposed regions were identical in >90% of the global *N.*

gonorrhoeae population (Fig. 6 and 7). Notable exceptions were PorB.IA, TbpA, and TbpB, which were highly polymorphic. However, our analysis revealed conserved surface loops of TbpA, associated with transferrin binding, that could be presented to the immune system separately from the rest of the protein by use of a delivery system, for instance. Considering the high degree of conservation, the distribution among *N. gonorrhoeae* strains globally, or the low-frequency sequence polymorphisms in surface loops suggests that ACP, AniA, BamA, BamE, MtrE, NspA, NGO0778, NGO1251, NGO1985, OpcA, PldA, Slam2, and ZnuD are promising candidates for a gonorrhea vaccine. Together, these results can inform gonorrhea vaccine development, including structural vaccinology efforts, by identifying conserved antigens, highlighting regions of conservation, and mapping polymorphic sites. Change in an antigen epitope landscape, however, is one of the hindrances in design of a successful vaccine. Important additional factors include bacterial escape from immune surveillance by reducing an antigen's prevalence/expression on the cell surface and intrinsic and extrinsic host-related, environmental, behavioral, nutritional, and vaccine administration factors (1, 86, 87).

MATERIALS AND METHODS

Allele mining, SNP/SAAP, and phylogenetic analyses. Nucleotide or amino acid sequences for all antigens were downloaded from the *Neisseria* PubMLST database as previously described (41), between January and May 2020. Alleles annotated as atypical or with premature stop codons were excluded from analysis. Nucleotide sequences were aligned in Geneious Prime 2020.1.2 using the Clustal Omega algorithm. The FA1090 allele was set as the reference sequence, automated SNP/SAAP annotation was performed in Geneious, and the results were exported to Excel. The FA1090 LbpA allele possesses a premature stop codon, so the most common allele (16) was used as the reference sequence for this locus (see Data Set S1 in the supplemental material). For phylogenetic and polymorphism mapping analyses, nucleotide sequences were translated and aligned by Muscle in Geneious. Neighbor-joining trees were assembled from aligned amino acid sequence data for each antigen by using the Jukes-Cantor distance model (88) in Geneious.

Allele proportion analysis. Using data from the two-field breakdown table from PubMLST, the allele number and the number of isolates associated with each allele were imported into GraphPad Prism v8.4.3 for Mac. "Parts of whole" graphs were generated automatically.

Polymorphism mapping. The applicable crystal structure sequence was aligned against translated sequences from all antigen alleles using Muscle in Geneious. The crystal structures of the following proteins were used in this analysis: ACP (PDB accession no. 6GQ4) (81), AniA (5TB7) (89), BamA (4K3B) (90), BamE (5WAM) (91), MtrE (4MT0) (92), NspA (1P4T) (93), OpcA (1K24) (94), PilQ (4AV2) (78), PorB.IA (4AUI) (95), TbpA (3V8X) (80), TbpB (3V8U) (80), and ZnuD (4RVW) (96).

The number of isolates associated with each polymorphic site in the alignment was calculated and divided by the total number of isolates with data for each antigen to determine the proportion of the population that possessed each polymorphism (number of isolates with polymorphism/total number of *N. gonorrhoeae* isolates). The polymorphisms associated with >1% of isolates were mapped to structural data using PyMOL (https://pymol.org), as previously described (41).

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.
DATA SET S1, XLSX file, 0.6 MB.
FIG S1, PDF file, 0.1 MB.
FIG S2, PDF file, 0.1 MB.
FIG S3, PDF file, 0.2 MB.
FIG S4, PDF file, 0.3 MB.
FIG S5, PDF file, 0.3 MB.
FIG S6, PDF file, 0.4 MB.
FIG S7, PDF file, 0.7 MB.
FIG S8, PDF file, 0.5 MB.
TABLE S1, PDF file, 0.02 MB.

ACKNOWLEDGMENTS

We thank the reviewers for insightful and detailed comments about our work.

This work was supported by National Institute of Allergy and Infectious Diseases grants R01AI117235 and U19 AI144180 (5967 and 5973).

The content is solely the authors' responsibility and does not necessarily represent the official views of the NIH. The funding sources had no role in study design, in collection, analysis, or interpretation of data, in manuscript preparation, or in the decision to submit the manuscript for publication.

A.E.S. developed and designed the study. B.I.B., R.A.Z., and A.K.H. collected and analyzed data. B.I.B., A.K.H., and A.E.S. wrote the manuscript. All authors provided feedback on the manuscript and approved the final version to be submitted. We attest that we meet the International Committee of Medical Journal Editors (ICMJE) criteria for authorship.

REFERENCES

1. Servin-Blanco R, Zamora-Alvarado R, Gevorkian G, Manoutcharian K. 2016. Antigenic variability: obstacles on the road to vaccines against traditionally difficult targets. Hum Vaccin Immunother 12:2640–2648. https://doi.org/10.1080/21645515.2016.1191718.
2. Palmer GH, Bankhead T, Seifert HS. 2016. Antigenic variation in bacterial pathogens. Microbiol Spectr 4:10.1128/microbiolspec.VMBF-0005-2015. https://doi.org/10.1128/microbiolspec.VMBF-0005-2015.
3. Deitsch KW, Lukehart SA, Stringer JR. 2009. Common strategies for antigenic variation by bacterial, fungal and protozoan pathogens. Nat Rev Microbiol 7:493–503. https://doi.org/10.1038/nrmicro2145.
4. Zhang G, Meredith TC, Kahne D. 2013. On the essentiality of lipopolysaccharide to Gram-negative bacteria. Curr Opin Microbiol 16:779–785. https://doi.org/10.1016/j.mib.2013.09.007.
5. Stern A, Brown M, Nickel P, Meyer TF. 1986. Opacity genes in Neisseria gonorrhoeae: control of phase and antigenic variation. Cell 47:61–71. https://doi.org/10.1016/0092-8674(86)90366-1.
6. Swanson J, Kraus SJ, Gotschlich EC. 1971. Studies on gonococcus infection. I. Pili and zones of adhesion: their relation to gonococcal growth patterns. J Exp Med 134:886–906. https://doi.org/10.1084/jem.134.4.886.
7. Snyder LA, Saunders NJ, Shafer WM. 2001. A putatively phase variable gene (dca) required for natural competence in Neisseria gonorrhoeae but not Neisseria meningitidis is located within the division cell wall (dcw) gene cluster. J Bacteriol 183:1233–1241. https://doi.org/10.1128/JB.183.4.1233-1241.2001.
8. EBioMedicine. 2017. Gonococcal vaccine development: lessons from group B meningococcal vaccines. EBioMedicine 22:1. https://doi.org/10.1016/j.ebiom.2017.07.026.
9. Giuliani MM, Adu-Bobie J, Comanducci M, Arico B, Savino S, Santini L, Brunelli B, Bambini S, Biolchi A, Capecchi B, Cartocci E, Ciucchi L, Di Marcello F, Ferlicca F, Galli B, Luzzi E, Masignani V, Serruto D, Veggi D, Contorni M, Morandi M, Bartalesi A, Cinotti V, Mannucci D, Titta F, Ovidi E, Welsch JA, Granoff D, Rappuoli R, Pizza M. 2006. A universal vaccine for serogroup B meningococcus. Proc Natl Acad Sci U S A 103:10834–10839. https://doi.org/10.1073/pnas.0603940103.
10. Shirley M, Dhillon S. 2015. Bivalent rLP2086 vaccine (Trumenba(®)): a review in active immunization against invasive meningococcal group B disease in individuals aged 10-25 years. BioDrugs 29:353–361. https://doi.org/10.1007/s40259-015-0139-0.
11. Fletcher LD, Bernfield L, Barniak V, Farley JE, Howell A, Knauf M, Ooi P, Smith RP, Weise P, Wetherell M, Xie X, Zagursky R, Zhang Y, Zlotnick GW. 2004. Vaccine potential of the Neisseria meningitidis 2086 lipoprotein. Infect Immun 72:2088–2100. https://doi.org/10.1128/IAI.72.4.2088-2100.2004.
12. Jiang HQ, Hoiseth SK, Harris SL, McNeil LK, Zhu D, Tan C, Scott AA, Alexander K, Mason K, Miller L, DaSilva I, Mack M, Zhao XJ, Pride MW, Andrew L, Murphy E, Hagen M, French R, Arora A, Jones TR, Jansen KU, Zlotnick GW, Anderson AS. 2010. Broad vaccine coverage predicted for a bivalent recombinant factor H binding protein based vaccine to prevent serogroup B meningococcal disease. Vaccine 28:6086–6093. https://doi.org/10.1016/j.vaccine.2010.06.083.
13. Rice PA, Shafer WM, Ram S, Jerse AE. 2017. Neisseria gonorrhoeae: drug resistance, mouse models, and vaccine development. Annu Rev Microbiol 71:665–686. https://doi.org/10.1146/annurev-micro-090816-093530.
14. CDC. 2019. Sexually transmitted diseases surveillance 2018. https://www.cdc.gov/std/stats18/default.htm.
15. CDC. 2015. Gonococcal infections: 2015 STD treatment guidelines. https://www.cdc.gov/std/tg2015/default.htm.
16. Workowski KA. 2015. Centers for Disease Control and Prevention sexually transmitted diseases treatment guidelines. Clin Infect Dis 61(Suppl 8):S759–S762. https://doi.org/10.1093/cid/civ771.
17. Thorpe P, Cyr SS, Barbee L, Workowski KA, Bachmann LH, Pham C, Schlanger K, Torrone E, Weinstock H, Kersh EN. 2020. Update to CDC's treatment guidelines for gonococcal infection. MMWR Morb Mortal Wkly Rep 69:1911–1916. https://doi.org/10.15585/mmwr.mm6950a6.
18. Unemo M, Shafer WM. 2014. Antimicrobial resistance in Neisseria gonorrhoeae in the 21st century: past, evolution, and future. Clin Microbiol Rev 27:587–613. https://doi.org/10.1128/CMR.00010-14.
19. Marjuki H, Topaz N, Joseph SJ, Gernert KM, Kersh EN, Antimicrobial-Resistant Neisseria gonorrhoeae Working Group, Wang X. 2019. Genetic similarity of gonococcal homologs to meningococcal outer membrane proteins of serogroup B vaccine. mBio 10:e01668-19. https://doi.org/10.1128/mBio.01668-19.
20. WHO. 2012. Global action plan to control the spread and impact of antimicrobial resistance in Neisseria gonorrhoeae. WHO, Geneva, Switzerland.
21. Dolange V, Churchward CP, Christodoulides M, Snyder LAS. 2018. The growing threat of gonococcal blindness. Antibiotics (Basel) 7:59. https://doi.org/10.3390/antibiotics7030059.
22. Harrison WO, Hooper RR, Wiesner PJ, Campbell AF, Karney WW, Reynolds GH, Jones OG, Holmes KK. 1979. A trial of minocycline given after exposure to prevent gonorrhea. N Engl J Med 300:1074–1078. https://doi.org/10.1056/NEJM197905103001903.
23. McCormack WM, Stumacher RJ, Johnson K, Donner A. 1977. Clinical spectrum of gonococcal infection in women. Lancet 1:1182–1185. https://doi.org/10.1016/S0140-6736(77)92720-9.
24. Handsfield HH, Lipman TO, Harnisch JP, Tronca E, Holmes KK. 1974. Asymptomatic gonorrhea in men. Diagnosis, natural course, prevalence and significance. N Engl J Med 290:117–123. https://doi.org/10.1056/NEJM197401172900301.
25. Platt R, Rice PA, McCormack WM. 1983. Risk of acquiring gonorrhea and prevalence of abnormal adnexal findings among women recently exposed to gonorrhea. JAMA 250:3205–3209. https://doi.org/10.1001/jama.1983.03340230057031.
26. Gottlieb SL, Ndowa F, Hook EW, III, Deal C, Bachmann L, Abu-Raddad L, Chen XS, Jerse A, Low N, MacLennan CA, Petousis-Harris H, Seib KL, Unemo M, Vincent L, Giersing BK, Gonococcal Vaccine PPC Expert Advisory Group. 2020. Gonococcal vaccines: public health value and preferred product characteristics; report of a WHO global stakeholder consultation, January 2019. Vaccine 38:4362–4373. https://doi.org/10.1016/j.vaccine.2020.02.073.
27. WHO. 2018. Antibiotic-resistant gonorrhoea on the rise, new drugs needed. WHO, Geneva, Switzerland.
28. Rendtorff RC, Curran JW, Chandler RW, Wiser WL, Robinson H. 1974. Economic consequences of gonorrhea in women: experience from an urban hospital. Sexually Transmitted Diseases 1:40–47.
29. Kirkcaldy RD, Weston E, Segurado AC, Hughes G. 2019. Epidemiology of gonorrhoea: a global perspective. Sex Health 16:401–411. https://doi.org/10.1071/SH19061.
30. ECDC. 2012. Response plan to control and manage the threat of multidrug-resistant gonorrhoea in Europe. ECDC, Solna, Sweden.
31. El-Rami FE, Zielke RA, Wi T, Sikora AE, Unemo M. 2019. Quantitative proteomics of the 2016 WHO Neisseria gonorrhoeae reference strains surveys vaccine candidates and antimicrobial resistance determinants. Mol Cell Proteomics 18:127–150. https://doi.org/10.1074/mcp.RA118.001125.
32. Zielke RA, Wierzbicki IH, Weber JV, Gafken PR, Sikora AE. 2014. Quantitative proteomics of the Neisseria gonorrhoeae cell envelope and membrane

vesicles for the discovery of potential therapeutic targets. Mol Cell Proteomics 13:1299–1317. https://doi.org/10.1074/mcp.M113.029538.
33. Zielke RA, Wierzbicki IH, Baarda BI, Gafken PR, Soge OO, Holmes KK, Jerse AE, Unemo M, Sikora AE. 2016. Proteomics-driven antigen discovery for development of vaccines against gonorrhea. Mol Cell Proteomics 15:2338–2355. https://doi.org/10.1074/mcp.M116.058800.
34. Phillips NJ, Steichen CT, Schilling B, Post DM, Niles RK, Bair TB, Falsetta ML, Apicella MA, Gibson BW. 2012. Proteomic analysis of Neisseria gonorrhoeae biofilms shows shift to anaerobic respiration and changes in nutrient transport and outermembrane proteins. PLoS One 7:e38303. https://doi.org/10.1371/journal.pone.0038303.
35. McClure R, Nudel K, Massari P, Tjaden B, Su X, Rice PA, Genco CA. 2015. The gonococcal transcriptome during infection of the lower genital tract in women. PLoS One 10:e0133982. https://doi.org/10.1371/journal.pone.0133982.
36. Baarda BI, Martinez FG, Sikora AE. 2018. Proteomics, bioinformatics and structure-function antigen mining for gonorrhea vaccines. Front Immunol 9:2793. https://doi.org/10.3389/fimmu.2018.02793.
37. Vincent LR, Jerse AE. 2019. Biological feasibility and importance of a gonorrhea vaccine for global public health. Vaccine 37:7419–7426. https://doi.org/10.1016/j.vaccine.2018.02.081.
38. Sikora AE, Gomez C, Le Van A, Baarda BI, Darnell S, Martinez FG, Zielke RA, Bonventre JA, Jerse AE. 2020. A novel gonorrhea vaccine composed of MetQ lipoprotein formulated with CpG shortens experimental murine infection. Vaccine 38:8175–8184. https://doi.org/10.1016/j.vaccine.2020.10.077.
39. Leduc I, Connolly KL, Begum A, Underwood K, Darnell S, Shafer WM, Balthazar JT, Macintyre AN, Sempowski GD, Duncan JA, Little MB, Rahman N, Garges EC, Jerse AE. 2020. The serogroup B meningococcal outer membrane vesicle-based vaccine 4CMenB induces cross-species protection against Neisseria gonorrhoeae. PLoS Pathog 16:e1008602. https://doi.org/10.1371/journal.ppat.1008602.
40. Jolley KA, Bray JE, Maiden MCJ. 2018. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. Wellcome Open Res 3:124. https://doi.org/10.12688/wellcomeopenres.14826.1.
41. Baarda BI, Zielke RA, Nicholas RA, Sikora AE. 2018. PubMLST for antigen allele mining to inform development of gonorrhea protein-based vaccines. Front Microbiol 9:2971. https://doi.org/10.3389/fmicb.2018.02971.
42. Baarda BI, Zielke RA, Le Van A, Jerse AE, Sikora AE. 2019. Neisseria gonorrhoeae MlaA influences gonococcal virulence and membrane vesicle production. PLoS Pathog 15:e1007385. https://doi.org/10.1371/journal.ppat.1007385.
43. Baarda BI, Sikora AE. 2019. Phenotypic microarray screening of Neisseria gonorrhoeae in chemically defined liquid medium. Methods Mol Biol 1997:207–216. https://doi.org/10.1007/978-1-4939-9496-0_13.
44. van der Woude MW, Baumler AJ. 2004. Phase and antigenic variation in bacteria. Clin Microbiol Rev 17:581–611. https://doi.org/10.1128/CMR.17.3.581-611.2004.
45. Unemo M, Golparian D, Sanchez-Buso L, Grad Y, Jacobsson S, Ohnishi M, Lahra MM, Limnios A, Sikora AE, Wi T, Harris SR. 2016. The novel 2016 WHO Neisseria gonorrhoeae reference strains for global quality assurance of laboratory investigations: phenotypic, genetic and reference genome characterization. J Antimicrob Chemother 71:3096–3108. https://doi.org/10.1093/jac/dkw288.
46. Baarda BI, Emerson S, Proteau PJ, Sikora AE. 2017. Deciphering the function of new gonococcal vaccine antigens using phenotypic microarrays. J Bacteriol 199:e00037-17. https://doi.org/10.1128/JB.00037-17.
47. Mulks MH, Knapp JS. 1987. Immunoglobulin A1 protease types of Neisseria gonorrhoeae and their relationship to auxotype and serovar. Infect Immun 55:931–936. https://doi.org/10.1128/IAI.55.4.931-936.1987.
48. Plaut AG, Gilbert JV, Artenstein MS, Capra JD. 1975. Neisseria gonorrhoeae and Neisseria meningitidis: extracellular enzyme cleaves human immunoglobulin A. Science 190:1103–1105. https://doi.org/10.1126/science.810892.
49. Pohlner J, Halter R, Meyer TF. 1987. Neisseria gonorrhoeae IgA protease. Secretion and implications for pathogenesis. Antonie Van Leeuwenhoek 53:479–484. https://doi.org/10.1007/BF00415506.
50. Pohlner J, Halter R, Beyreuther K, Meyer TF. 1987. Gene structure and extracellular secretion of Neisseria gonorrhoeae IgA protease. Nature 325:458–462. https://doi.org/10.1038/325458a0.
51. Halter R, Pohlner J, Meyer TF. 1984. IgA protease of Neisseria gonorrhoeae: isolation and characterization of the gene and its extracellular product. EMBO J 3:1595–1601. https://doi.org/10.1002/j.1460-2075.1984.tb02016.x.
52. Plaut AG, Gilbert JV, Wistar R. Jr, 1977. Loss of antibody activity in human immunoglobulin A exposed extracellular immunoglobulin A proteases of Neisseria gonorrhoeae and Streptococcus sanguis. Infect Immun 17:130–135. https://doi.org/10.1128/IAI.17.1.130-135.1977.
53. Senior BW, Stewart WW, Galloway C, Kerr MA. 2001. Cleavage of the hormone human chorionic gonadotropin, by the type 1 IgA1 protease of Neisseria gonorrhoeae, and its implications. J Infect Dis 184:922–925. https://doi.org/10.1086/323397.
54. Lin L, Ayala P, Larson J, Mulks M, Fukuda M, Carlsson SR, Enns C, So M. 1997. The Neisseria type 2 IgA1 protease cleaves LAMP1 and promotes survival of bacteria within epithelial cells. Mol Microbiol 24:1083–1094. https://doi.org/10.1046/j.1365-2958.1997.4191776.x.
55. Shoberg RJ, Mulks MH. 1991. Proteolysis of bacterial membrane proteins by Neisseria gonorrhoeae type 2 immunoglobulin A1 protease. Infect Immun 59:2535–2541. https://doi.org/10.1128/IAI.59.8.2535-2541.1991.
56. Simpson DA, Hausinger RP, Mulks MH. 1988. Purification, characterization, and comparison of the immunoglobulin A1 proteases of Neisseria gonorrhoeae. J Bacteriol 170:1866–1873. https://doi.org/10.1128/jb.170.4.1866-1873.1988.
57. Paruchuri DK, Seifert HS, Ajioka RS, Karlsson KA, So M. 1990. Identification and characterization of a Neisseria gonorrhoeae gene encoding a glycolipid-binding adhesin. Proc Natl Acad Sci U S A 87:333–337. https://doi.org/10.1073/pnas.87.1.333.
58. Jamet A, Jousset AB, Euphrasie D, Mukorako P, Boucharlat A, Ducousso A, Charbit A, Nassif X. 2015. A new family of secreted toxins in pathogenic Neisseria species. PLoS Pathog 11:e1004592. https://doi.org/10.1371/journal.ppat.1004592.
59. Haines KA, Yeh L, Blake MS, Cristello P, Korchak H, Weissmann G. 1988. Protein I, a translocatable ion channel from Neisseria gonorrhoeae, selectively inhibits exocytosis from human neutrophils without inhibiting O_2^- generation. J Biol Chem 263:945–951. https://doi.org/10.1016/S0021-9258(19)35444-4.
60. Unemo M, Olcen P, Berglund T, Albert J, Fredlund H. 2002. Molecular epidemiology of Neisseria gonorrhoeae: sequence analysis of the porB gene confirms presence of two circulating strains. J Clin Microbiol 40:3741–3749. https://doi.org/10.1128/jcm.40.10.3741-3749.2002.
61. Brunham RC, Plummer F, Slaney L, Rand F, DeWitt W. 1985. Correlation of auxotype and protein I type with expression of disease due to Neisseria gonorrhoeae. J Infect Dis 152:339–343. https://doi.org/10.1093/infdis/152.2.339.
62. Cannon JG, Buchanan TM, Sparling PF. 1983. Confirmation of association of protein I serotype of Neisseria gonorrhoeae with ability to cause disseminated infection. Infect Immun 40:816–819. https://doi.org/10.1128/IAI.40.2.816-819.1983.
63. Ram S, Sharma AK, Simpson SD, Gulati S, McQuillen DP, Pangburn MK, Rice PA. 1998. A novel sialic acid binding site on factor H mediates serum resistance of sialylated Neisseria gonorrhoeae. J Exp Med 187:743–752. https://doi.org/10.1084/jem.187.5.743.
64. Ram S, McQuillen DP, Gulati S, Elkins C, Pangburn MK, Rice PA. 1998. Binding of complement factor H to loop 5 of porin protein 1A: a molecular mechanism of serum resistance of nonsialylated Neisseria gonorrhoeae. J Exp Med 188:671–680. https://doi.org/10.1084/jem.188.4.671.
65. Ram S, Cullinane M, Blom AM, Gulati S, McQuillen DP, Monks BG, O'Connell C, Boden R, Elkins C, Pangburn MK, Dahlback B, Rice PA. 2001. Binding of C4b-binding protein to porin: a molecular mechanism of serum resistance of Neisseria gonorrhoeae. J Exp Med 193:281–295. https://doi.org/10.1084/jem.193.3.281.
66. Chen A, Seifert HS. 2013. Structure-function studies of the Neisseria gonorrhoeae major outer membrane porin. Infect Immun 81:4383–4391. https://doi.org/10.1128/IAI.00367-13.
67. Pajón R, Chinea G, Marrero E, Gonzalez D, Guillén G. 1997. Sequence analysis of the structural tbpA gene: protein topology and variable regions within neisserial receptors for transferrin iron acquisition. Microb Pathog 23:71–84. https://doi.org/10.1006/mpat.1997.0136.
68. Cornelissen CN, Biswas GD, Tsai J, Paruchuri DK, Thompson SA, Sparling PF. 1992. Gonococcal transferrin-binding protein 1 is required for transferrin utilization and is homologous to TonB-dependent outer membrane receptors. J Bacteriol 174:5788–5797. https://doi.org/10.1128/jb.174.18.5788-5797.1992.
69. Anderson JE, Sparling PF, Cornelissen CN. 1994. Gonococcal transferrin-binding protein 2 facilitates but is not essential for transferrin utilization. J Bacteriol 176:3162–3170. https://doi.org/10.1128/JB.176.11.3162-3170.1994.
70. Rokbi B, Renauld-Mongenie G, Mignon M, Danve B, Poncet D, Chabanel C,

Caugant DA, Quentin-Millet MJ. 2000. Allelic diversity of the two transferrin binding protein B gene isotypes among a collection of Neisseria meningitidis strains representative of serogroup B disease: implication for the composition of a recombinant TbpB-based vaccine. Infect Immun 68:4938–4947. https://doi.org/10.1128/iai.68.9.4938-4947.2000.

71. Harrison OB, Maiden MC, Rokbi B. 2008. Distribution of transferrin binding protein B gene (tbpB) variants among Neisseria species. BMC Microbiol 8:66. https://doi.org/10.1186/1471-2180-8-66.
72. Kuroda Y, Tsuyuguchi T, Sakai Y, K CS, Ishihara T, Yamaguchi T, Saisho H, Yokosuka O. 2010. Long-term follow-up evaluation for more than 10 years after endoscopic treatment for postoperative bile duct strictures. Surg Endosc 24:834–840. https://doi.org/10.1007/s00464-009-0673-2.
73. Zhu P, Morelli G, Achtman M. 1999. The opcA and (psi)opcB regions in Neisseria: genes, pseudogenes, deletions, insertion elements and DNA islands. Mol Microbiol 33:635–650. https://doi.org/10.1046/j.1365-2958.1999.01514.x.
74. Jerse AE, Sharma ND, Simms AN, Crow ET, Snyder LA, Shafer WM. 2003. A gonococcal efflux pump system enhances bacterial survival in a female mouse model of genital tract infection. Infect Immun 71:5576–5582. https://doi.org/10.1128/iai.71.10.5576-5582.2003.
75. Lee EH, Shafer WM. 1999. The farAB-encoded efflux pump mediates resistance of gonococci to long-chained antibacterial fatty acids. Mol Microbiol 33:839–845. https://doi.org/10.1046/j.1365-2958.1999.01530.x.
76. Lewis LA, Rice PA, Ram S. 2019. Role of gonococcal neisserial surface protein A (NspA) in serum resistance and comparison of its factor H binding properties with those of its meningococcal counterpart. Infect Immun 87:e00658-18. https://doi.org/10.1128/IAI.00658-18.
77. Helm RA, Barnhart MM, Seifert HS. 2007. pilQ Missense mutations have diverse effects on PilQ multimer formation, piliation, and pilus function in Neisseria gonorrhoeae. J Bacteriol 189:3198–3207. https://doi.org/10.1128/JB.01833-06.
78. Berry JL, Phelan MM, Collins RF, Adomavicius T, Tonjum T, Frye SA, Bird L, Owens R, Ford RC, Lian LY, Derrick JP. 2012. Structure and assembly of a trans-periplasmic channel for type IV pili in Neisseria meningitidis. PLoS Pathog 8:e1002923. https://doi.org/10.1371/journal.ppat.1002923.
79. Noinaj N, Buchanan SK, Cornelissen CN. 2012. The transferrin-iron import system from pathogenic Neisseria species. Mol Microbiol 86:246–257. https://doi.org/10.1111/mmi.12002.
80. Noinaj N, Easley NC, Oke M, Mizuno N, Gumbart J, Boura E, Steere AN, Zak O, Aisen P, Tajkhorshid E, Evans RW, Gorringe AR, Mason AB, Steven AC, Buchanan SK. 2012. Structural basis for iron piracy by pathogenic Neisseria. Nature 483:53–58. https://doi.org/10.1038/nature10823.
81. Almonacid-Mendoza HL, Humbert MV, Dijokaite A, Cleary DW, Soo Y, Hung MC, Orr CM, Machelett MM, Tews I, Christodoulides M. 2018. Structure of the recombinant Neisseria gonorrhoeae adhesin complex protein (rNg-ACP) and generation of murine antibodies with bactericidal activity against gonococci. mSphere 3:e00331-18. https://doi.org/10.1128/mSphere.00331-18.
82. Hollingshead S, Jongerius I, Exley RM, Johnson S, Lea SM, Tang CM. 2018. Structure-based design of chimeric antigens for multivalent protein vaccines. Nat Commun 9:1051. https://doi.org/10.1038/s41467-018-03146-7.
83. Gu H, Liao Y, Zhang J, Wang Y, Liu Z, Cheng P, Wang X, Zou Q, Gu J. 2018. Rational design and evaluation of an artificial Escherichia coli K1 protein vaccine candidate based on the structure of OmpA. Front Cell Infect Microbiol 8:172. https://doi.org/10.3389/fcimb.2018.00172.
84. Xu L, He D, Yang L, Li Z, Ye X, Yu H, Zhao H, Li S, Yuan L, Qian H, Que Y, Kuo Shih JW, Zhu H, Li Y, Cheng T, Xia N. 2015. A broadly cross-protective vaccine presenting the neighboring epitopes within the VP1 GH loop and VP2 EF loop of enterovirus 71. Sci Rep 5:12973. https://doi.org/10.1038/srep12973.
85. Cardozo T, Wang S, Jiang X, Kong XP, Hioe C, Krachmarov C. 2014. Vaccine focusing to cross-subtype HIV-1 gp120 variable loop epitopes. Vaccine 32:4916–4924. https://doi.org/10.1016/j.vaccine.2014.07.026.
86. Rollenhagen C, Sörensen M, Rizos K, Hurvitz R, Bumann D. 2004. Antigen selection based on expression levels during infection facilitates vaccine development for an intracellular pathogen. Proc Natl Acad Sci U S A 101:8739–8744. https://doi.org/10.1073/pnas.0401283101.
87. Zimmermann P, Curtis N. 2019. Factors that influence the immune response to vaccination. Clin Microbiol Rev 32:e00084-18. https://doi.org/10.1128/CMR.00084-18.
88. Jukes TH, Cantor CR. 1969. Evolution of protein molecules, p 21–131. *In* Munro HN (ed), Mammalian protein metabolism. Academic Press, New York, NY.
89. Sikora AE, Mills RH, Weber JV, Hamza A, Passow BW, Romaine A, Williamson ZA, Reed RW, Zielke RA, Korotkov KV. 2017. Peptide inhibitors targeting the Neisseria gonorrhoeae pivotal anaerobic respiration factor AniA. Antimicrob Agents Chemother 61:e00186-17. https://doi.org/10.1128/AAC.00186-17.
90. Noinaj N, Kuszak AJ, Gumbart JC, Lukacik P, Chang H, Easley NC, Lithgow T, Buchanan SK. 2013. Structural insight into the biogenesis of beta-barrel membrane proteins. Nature 501:385–390. https://doi.org/10.1038/nature12521.
91. Sikora AE, Wierzbicki IH, Zielke RA, Ryner RF, Korotkov KV, Buchanan SK, Noinaj N. 2018. Structural and functional insights into the role of BamD and BamE within the beta-barrel assembly machinery in Neisseria gonorrhoeae. J Biol Chem 293:1106–1119. https://doi.org/10.1074/jbc.RA117.000437.
92. Lei HT, Chou TH, Su CC, Bolla JR, Kumar N, Radhakrishnan A, Long F, Delmar JA, Do SV, Rajashankar KR, Shafer WM, Yu EW. 2014. Crystal structure of the open state of the Neisseria gonorrhoeae MtrE outer membrane channel. PLoS One 9:e97475. https://doi.org/10.1371/journal.pone.0097475.
93. Vandeputte-Rutten L, Bos MP, Tommassen J, Gros P. 2003. Crystal structure of Neisserial surface protein A (NspA), a conserved outer membrane protein with vaccine potential. J Biol Chem 278:24825–24830. https://doi.org/10.1074/jbc.M302803200.
94. Prince SM, Achtman M, Derrick JP. 2002. Crystal structure of the OpcA integral membrane adhesin from Neisseria meningitidis. Proc Natl Acad Sci U S A 99:3417–3421. https://doi.org/10.1073/pnas.062630899.
95. Zeth K, Kozjak-Pavlovic V, Faulstich M, Fraunholz M, Hurwitz R, Kepp O, Rudel T. 2013. Structure and function of the PorB porin from disseminating Neisseria gonorrhoeae. Biochem J 449:631–642. https://doi.org/10.1042/BJ20121025.
96. Calmettes C, Ing C, Buckwalter CM, El Bakkouri M, Chieh-Lin Lai C, Pogoutse A, Gray-Owen SD, Pomes R, Moraes TF. 2015. The molecular mechanism of Zinc acquisition by the neisserial outer-membrane transporter ZnuD. Nat Commun 6:7996. https://doi.org/10.1038/ncomms8996.

HHS Public Access
Author manuscript
Brainlesion. Author manuscript; available in PMC 2019 September 03.

Published in final edited form as:
Brainlesion. 2019 ; 11383: 239–250. doi:10.1007/978-3-030-11723-8_24.

Multi-stage Association Analysis of Glioblastoma Gene Expressions with Texture and Spatial Patterns

Samar S. M. Elsheikh[1], **Spyridon Bakas**[2,3], **Nicola J. Mulder**[1], **Emile R. Chimusa**[4], **Christos Davatzikos**[2,3], **Alessandro Crimi**[5,6]

[1]Computational Biology Group, Department of Integrative Biomedical Sciences, Institute of Infectious Disease and Molecular Medicine, Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa [2]Center for Biomedical Image Computing and Analytics (CBICA), Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA [3]Department of Radiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA [4]Division of Human Genetics, Institute of Infectious Disease and Molecular Medicine, Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa [5]University Hospital of Zürich, Zürich, Switzerland [6]African Institute for Mathematical Sciences, Biriwa, Ghana

Abstract

Glioblastoma is the most aggressive malignant primary brain tumor with a poor prognosis. Glioblastoma heterogeneous neuroimaging, pathologic, and molecular features provide opportunities for subclassification, prognostication, and the development of targeted therapies. Magnetic resonance imaging has the capability of quantifying specific phenotypic imaging features of these tumors. Additional insight into disease mechanism can be gained by exploring genetics foundations. Here, we use the gene expressions to evaluate the associations with various quantitative imaging phenomic features extracted from magnetic resonance imaging. We highlight a novel correlation by carrying out multi-stage genomewide association tests at the gene-level through a non-parametric correlation framework that allows testing multiple hypotheses about the integrated relationship of imaging phenotype-genotype more efficiently and less expensive computationally. Our result showed several novel genes previously associated with glioblastoma and other types of cancers, as the LRRC46 (chromosome 17), EPGN (chromosome 4) and TUBA1C (chromosome 12), all associated with our radiographic tumor features.

Keywords

Glioblastoma; Gene expression; Brain tumor; Radiomics; Radiogenomics

1 Introduction

Gliomas are the most common type of primary adult brain tumors that arise from glial cells. Gliomas have a very heterogeneous landscape, and they can be classified according to their

---

Samar.salah119@gmail.com.

Author Manuscript

grade into low-grade glioma, anaplastic glioma, and glioblastoma. The most common and aggressive type of glioma in adults is glioblastoma (GBM), which gives to the affected patient average survival time of only 10 to 18 months. The known molecular classification of GBM into classical, mesenchymal, neural and proneural subtypes is relatively accepted to be related to the expression of EGFR, NF1 and PDGFRA/IDH1 genes [1].

Imaging, specifically magnetic resonance imaging (MRI), can offer data towards promising biomarkers reflecting underlying tumor pathology and biological function. If imaging phenotypes of GBM obtained from routine clinical MRI studies can be associated with specific gene expression signatures, quantitative imaging phenotypes will serve as non-invasive surrogates for cancer genomic events and provide valuable information as to the diagnosis, prognosis, and optimal treatment.

Several radiogenomic studies have been carried out for many diseases [8–16]. For instance for schizophrenia pairs of SNP/Gene and MRI features have been mapped by using PLINK [8], and Parallel-ICA showed promising results [9]. Batmanghelich et al. [10] proposed a Bayesian framework to relate imaging and genetic data to phenotypes exploiting connection among these data modalities simultaneously in Alzheimer. Recently, correlations of connectomic features have been related to genes which are known to be related to Alzheimer progression [11]. In contrast to Alzheimer's disease and schizophrenia, glioma lesions are generally not spread all over the brain, and local features from MRI can be used. An imaging-genomic analysis study [12], performed by using the tumor volume in T2-weighted FLuid-Attenuated Inversion Recovery (T2-FLAIR) images and large-scale genetic and micro-RNA expression probes demonstrated the potential for molecular subtyping and showed that the high median expression of POSTN gene results in a significant decrease in survival, and for that they used ANOVA and Tukey-Kramer test. Other studies [13,14] showed correlations between image feature annotations and expression of genes with glioma molecular subtypes [1]. Specifically, Gutman et al. [13] found a significant association between contrast-enhanced tumor and these molecular subtypes [1], where proneural type expressed by PDGFRA/IDH1 gene showed low levels of contrast enhancement, and the classical type (i.e., primarily described by EGFR amplification) correlates with the increased percentage of contrast enhancement. The study used sher exact statistics.

Recent population-based studies have assessed the anatomical location of GBM in relation to distinct clinically-relevant molecular characteristics, and have identified the spatial distribution of the tumors being descriptive of their molecular status [14,17–22]. Furthermore, the emerging research direction of radiomics has shown promise that texture analysis of the various tumor sub-regions in radiographic imaging can also be informative of the tumor's molecular characterization [23–25]. Furthermore, using MRI features for GBM lesions, including texture and shape features, Haruka et al. proposed a classification imaging method and found three clusters of GBM patients [35]. In their method, they integrate copy number and gene expression data to estimate the molecular pathway activity and show that the three clusters reveal not only different molecular characteristics but also different survival probabilities.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

The purpose of this paper is to identify significant associations between gene expressions, across the whole genome, and quantitative imaging phenomic features extracted from multi-modal MRI brain scans of patients diagnosed with de novo primary GBM. In line with the pre-mentioned studies, here we focus on evaluating the spatial location and texture features of GBM and investigate their associations with gene expressions.

2 Materials and Methods

2.1 Data

For the quantitative association analysis conducted here, we utilized a retrospective cohort of 135 de novo primary GBM patients from the TCGA-GBM collection [6], with available pre-operative multi-modal MRI scans in The Cancer Imaging Archive (TCIA) [7] and corresponding molecular characterization in The Cancer Genome Atlas (TGCA). The multi-modal MRI data we utilized comprise native (T1) and post-contrast T1-weighted (T1Gd), T2-weighted (T2), and T2-FLAIR modalities. The TCGA-GBM subset of 135 patients were identified by Bakas et al. [4] as brain scans without any surgically-imposed cavity, and their co-registered and skull-stripped imaging were provided in the TCIA Analysis Results together with expert manually annotated segmentation labels for the various histologically-distinct tumor sub-regions, i.e. enhancing tumor (ET), non-enhancing tumor (NET), peritumoral edematous/invaded tissue (ED) (Fig. 1) [4,5]. The total sample size of GBM patients reduced to 88 after evaluating patients that had available imaging [6] and corresponding gene expressions. In total, we assessed expression energies for 17815 genes, 11 distinct descriptors of tumor spatial location (Fig. 2), and 517 radiomic/texture features (Fig. 2) for each patient's brain tumor scan [2,4,5].

2.2 Quantitative Imaging Phenomic Features

Radiomic/Texture Features.—We extracted an extensive panel of quantitative texture features, volumetrically (in 3D), for each tumor sub-region as provided by the expert annotations, across all available modalities. Specifically, the texture features we evaluated (i) capture global characteristics (i.e., variance, skewness, kurtosis) of each sub-region's intensity distribution on each modality, and (ii) include features based on Gray-Level Co-occurrence Matrix (GLCM) [26] (Fig. 2), Gray-Level Run-Length Matrix (GLRLM) [27–30], Gray-Level Size Zone Matrix (GLSZM) [28–30], and Neighborhood Gray-Tone Difference Matrix (NGTDM) [31].

Spatial Distribution Patterns.—Beyond texture features, we collected discrete spatial information about the anatomical location of each tumor on each brain scan (Fig. 2). To obtain these spatial distribution patterns we registered all brain tumor scans in a standardized healthy atlas space using an iterative Expectation Maximization framework [3], while incorporating a biophysical tumor growth model (based on a reaction-diffusion-advection model [32–34]) to account for tumor mass effects in the brain parenchyma. We then retrieved the spatial distribution of each tumor according to the discretized anatomical locations of the (i) specific lobes (i.e., frontal, temporal, parietal, occipital), (ii) insula, (iii) basal ganglia, (iv) fornix, (v) cerebellum, and (vi) brain stem. In addition, we also included

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

as distinct features the distances of (i) the tumor core (defined as the union of ET and NET), and (ii) the ED, from the ventricles.

To produce these quantitative features we have utilized GLISTRboost. Specifically, in the process to produce segmentations of the various tumor subregions, the generative part [37] of GLISTRboost, following an Expectation Maximization framework registers a healthy population probabilistic atlas to glioma patients' brain scans while incorporating a biophysical glioma growth model to account for mass effects. Then, after converting the predicted segmentation in the healthy atlas space, the percentage of the tumor core (i.e., enhancing and non-enhancing tumor) is calculated on each of the brain lobes in this healthy atlas.

2.3 Data Analysis

Initially, we combined the two types of data (imaging - genetics) using the patient ID as a primary column. As a first stage, we used the gene expressions and the spatial distribution patterns to perform a non-parametric test of association. To assess the associations, we computed the Spearman correlation coefficient (r_s) between the gene expressions, individually, as a with each of the spatial distribution patterns described in Sect. 2.2. We then assessed the significant of the correlation coefficient by calculating the p-values as described below.

For each quantitative feature and each gene, We obtained the p-value associated with Spearman correlation coefficient test statistic. That is, the p-value of the correlation between a single gene expression with a single feature of the tumor's location in the brain. The Spearman correlation coefficient model for a given feature (y) and given gene expression (x) is;

$$r_s = 1 - \frac{6\sum_{i=1}^{i=N} d_i^2}{N\left(N^2 - 1\right)} \quad (1)$$

Where d_i is the difference between the ranks of x_i and y_i, and N is 88; representing the number of GBM patients [38]. r_s can take any real value between +1 and −1; +1 represents a strong positive association, −1 means a perfect negative association and 0 indicates no association between the ranks of x and y. Our hypothesis of interest is:

H_0: There is no association between the gene expression and the tumor's feature under study

vs

H_a: There is an association between the gene expression and the tumor's feature under study, alternatively:

H_0: $\mathrm{r_s} = 0$ $vs\ H_0 : r_s \neq 0$

To determine the significance of r_s, one can use the t test statistic defined as

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$t_c = r_s\sqrt{\frac{n-2}{1-r_s^2}}, \quad (2)$$

t_c follows approximately the Student's t distribution with a $N2$ degrees of freedom under the null hypothesis [38]. At a certain significance level, the calculated value of t_c can be compared to the table value obtained from the Student's t distribution (as described previously). The significance of r_s can also be determined using the p-value, which is simply the integration, or the area under the curve from t_c to infinity.

Briefly, in this first stage, the association test was initially conducted to six features of the tumor location (Sect. 2.2). More specifically, for each gene, we computed six p-values, then considered only the minimum p-value at each gene (see Fig. 3 for the analysis workflow). The latter is referred to as meta-analysis in Fig. 3 (step(c)). All results reported in Sect. 3 use the summary statistics of the meta-analysis. Moreover, out of the all the association results, we excluded all the genes with p-values greater than or equal 0.05. Here we meant to exclude the genes that have very low (and not significant) association with the spatial pattern, which we believe is an important phenotype. This step is referred to as (d) in Fig. 3. In the second stage, we proceeded with all the genes with p-value less than 0.05, excluding the least significant genes, and we carried the same analysis as in the first stage but using the radiomic features (Sect. 2.2. Table 1 shows the thresholds at both 5% and 10% significance level), along with the number of genes used and remained in each stage.

It is worth mentioning that, out of the total number of genes, we were able to annotate 15009 genes and assign them to their defined physical locations in the DNA. We carried on the first stage of the analysis using those genes (Table 1).

3 Results

The incidence of tumors specific for region is summarized in Table 2. The Manhattan plot for the p-values obtained from the meta-analysis is illustrated in Fig. 4. The plot shows two horizontal lines which associate with the thresholds of 5% significance level (top line), and 10% significance level (bottom line), after correcting for multiple comparisons. The x-axis is the physical position of genes in the DNA, and the y-axis is the negative $log10$ of the p-values. Figure 4 also shows the qq-plot of all the genes used in the association analysis. Likewise, each dot corresponds to a p-value of a single gene and $-log10$ of the p-value is used instead. The qq-plot reported with each Manhattan plot, and it compares the observed distribution of p-values (y-axis) to the expected distribution (x-axis), for each gene tested, where the diagonal line is the null distribution.

Table 3 shows (only) the highest ten p-values and the corresponding genes of the first stage of the analysis. In this stage, non of the p-values was less than $3.3e^{-6}$ or $6.7e^{-6}$ (see Table 1); therefore, no gene was significantly associated with any of the features. Table 3 reports the gene symbol, its start and end position, the associated p-value and feature, and the chromosome.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

We then pruned the genes used in the previous stage to a smaller set, by removing the genes that have p-values less than 0.05. With the 5401 genes remaining, we took over the second stage and repeated the same analysis with the texture characteristics of the tumor. The Manhattan and qq-plot for the texture features are shown in Fig. 5, and Table 4 shows the top 10 significant genes. Total of significant genes in this stage is 37 (at 5% significance level).

4 Discussion

GBM is a fatal malignant disease that so far is incurable. The identification of genetic risk factors that affect the tumor characteristics improves our understanding of the underlying biological processes for GBM, and contribute to therapeutic discovery. In this study, we proposed a framework that allows quantifying the non-parametric correlations to test associations between gene expressions and different quantitative imaging phenomic characteristics of GBM. Our result has shown a high genetic enrichment through the Manhattan and qq-plots, especially for the texture features (Fig. 5).

Our results highlighted several genes that significantly associated with the tumor texture features, including *LRRC46, USP38, EPGN, TUBA1C, ZNF284, IPO8, MMP7, TLL2, TRIM55* and *UBAP1*, as the top ten significant genes (Table 4). However, there are, in total, 37 genes are significantly associated with the texture features (Fig. 5). EPGN expression associates significantly ($r_s = 0.501$, p-value = $6.542e{-}07$) with GLSZM LGZE in the T1Gd modality (Table 4). EPGN previously reported to be one of the top ten upregulated genes after EBLN1 silencing in oligodendroglia cells [39]. Moreover, the emergence of EPGN was marked in another study by Duhem-Tonnelle et al. in EGF ligands expression profile, between glioblastoma cell lines and biopsies [40]. Located at chromosome 4, USP38 ($r_s =$ -0.511, p-value = $3.648e{-}07$) [41]. Moreover, as it is illustrated in the Manhattan plot of the spatial features of the tumor (Fig. 4 and Table 3), no gene shows significant association with any of the location features. In addition to the latter, the number of GBM lesions in the cerebellum in clinical settings are quite rare [36], as also shown in our summary Table 2. Our study can give some insight into this rare type of GBM lesion. Nevertheless, the investigation excluding the patients having those lesions have to be repeated as a future work.

5 Conclusion

As the understanding of gliomagenesis grows, several medical imaging biomarkers and genetic variations can be identified, and new hypotheses can be formed. The hereby proposed genome-wide association framework aims at identifying differentially expressed genes that significantly correlate with various aspects of GBM. The identification of such genes may contribute to the development of targeted therapies that focus on the resistance mechanisms of individual patients.

Through the systematic testing of associations and shrinking of the number of genes at every stage, this pipeline facilitates the evaluation of various hypotheses and reduces the computational complexity. In future work, we plan to extend the study by integrating more

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

quantitative imaging phenomic tumor characteristics, inclusive of morphological, intensity, and volumetric descriptors, as well as parameters derived by biophysical tumor growth modeling.

Acknowledgement.

Research reported in this publication was partly supported by the National Institutes of Health (NIH) under award numbers NIH/NINDS:R01NS042645, NIH/NCI:U24CA189523, NIH/NCATS:UL1TR001878, the ITMAT of the University of Pennsylvania as well as by the Swedish International Development Cooperation Agency (SIDA) through the Organization for Women in Science for the Developing World (OWSD). Computations were performed using facilities provided by the University of Cape Town's ICTS High Performance Computing team: hpc.uct.ac.za.

References

1. Verhaak RGW, et al.: Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell 17(1), 98–110 (2010) [PubMed: 20129251]
2. Davatzikos C, et al.: Cancer imaging phenomics toolkit: quantitative imaging analytics for precision diagnostics and predictive modeling of clinical outcome. J. Med. Imaging 5(1), 011018 (2018)
3. Bakas S, et al.: GLISTRboost: combining multimodal MRI segmentation, registration, and biophysical tumor growth modeling with gradient boosting machines for glioma segmentation In: Crimi A, Menze B, Maier O, Reyes M, Handels H (eds.) BrainLes 2015. LNCS, vol. 9556, pp. 144–155. Springer, Cham (2016). 10.1007/978-3-319-30858-6_13
4. Bakas S, et al.: Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. Nat. Sci. Data 4, 170117 (2017)
5. Bakas S, et al.: Segmentation labels and radiomic features for the pre-operative scans of the TCGA-GBM collection The Cancer Imaging Archive (2017). 10.7937/K9/TCIA/.2017.KLXWJJ1Q
6. Scarpace L, et al.: Radiology data from the cancer genome atlas glioblastoma multiforme [TCGA-GBM] collection. Cancer Imaging Arch. 11, 4 (2016)
7. Clark K, et al.: The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. J. Digit. Imaging 26(6), 1045–1057 (2013) [PubMed: 23884657]
8. Stein JL, et al.: Voxelwise genome-wide association study (vGWAS). Neuroimage 53(3), 1160–1174 (2010) [PubMed: 20171287]
9. Liu J, et al.: Combining fMRI and SNP data to investigate connections between brain function and genetics using parallel ICA. Hum. Brain Mapp. 30(1), 241–255 (2009) [PubMed: 18072279]
10. Batmanghelich NK, Dalca AV, Sabuncu MR, Golland P: Joint modeling of imaging and genetics In: Gee JC, Joshi S, Pohl KM, Wells WM, Zöllei L (eds.) IPMI 2013. LNCS, vol. 7917, pp. 766–777. Springer, Heidelberg (2013). 10.1007/978-3-642-38868-2_64
11. Elsheikh S, et al.: Relating connectivity changes in brain networks to genetic information in Alzheimer patients In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). IEEE (2018)
12. Zinn PO, et al.: Radiogenomic mapping of edema/cellular invasion MRI phenotypes in glioblastoma multiforme. PLoS ONE 6(10), e25451 (2011) [PubMed: 21998659]
13. Gutman DA, et al.: MR imaging predictors of molecular profile and survival: multi-institutional study of the TCGA glioblastoma data set. Radiology 267(2), 560–569 (2013) [PubMed: 23392431]
14. Macyszyn L, et al.: Imaging patterns predict patient survival and molecular subtype in glioblastoma via machine learning techniques. Neuro-oncology 18(3), 417–425 (2015) [PubMed: 26188015]
15. Binder Z, et al.: Epidermal growth factor receptor extracellular domain mutations in glioblastoma present opportunities for clinical imaging and therapeutic development. Cancer Cell 34, 163–177 (2018) [PubMed: 29990498]

16. Bakas S, et al.: In vivo detection of EGFRvIII in glioblastoma via perfusion magnetic resonance imaging signature consistent with deep peritumoral in ltration: the φ-index. Clin. Cancer Res. 23, 4724–4734 (2017) [PubMed: 28428190]
17. Cancer Genome Atlas Research Network: Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. N. Engl. J. Med 372(26), 2481–2498 (2015) [PubMed: 26061751]
18. Ellingson BM, et al.: Probabilistic radiographic atlas of glioblastoma phenotypes. Am. J. Neuroradiol 34(3), 533–540 (2012) [PubMed: 22997168]
19. Ellingson BM: Radiogenomics and imaging phenotypes in glioblastoma: novel observations and correlation with molecular characteristics. Curr. Neurol. Neurosci. Rep 15(1), 506 (2015) [PubMed: 25410316]
20. Steed TC, et al.: Differential localization of glioblastoma subtype: implications on glioblastoma pathogenesis. Oncotarget 7(18), 24899 (2016) [PubMed: 27056901]
21. Bilello M, et al.: Population-based MRI atlases of spatial distribution are specific to patient and tumor characteristics in glioblastoma. NeuroImage: Clin. 12, 34–40 (2016) [PubMed: 27358767]
22. Akbari H, et al.: In vivo evaluation of EGFRvIII mutation in primary glioblastoma patients via complex multiparametric MRI signature. Neuro-Oncology 20(8), 1068–1079 (2018) [PubMed: 29617843]
23. Aerts HJWL: The potential of radiomic-based phenotyping in precision medicine: a review. JAMA Oncol. 2(12), 1636–1642 (2016) [PubMed: 27541161]
24. Lambin P, et al.: Radiomics: extracting more information from medical images using advanced feature analysis. Eur. J. Cancer 48(4), 441–446 (2012) [PubMed: 22257792]
25. Aerts HJWL, et al.: Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat. Commun 5, 4006 (2014) [PubMed: 24892406]
26. Haralick RM, et al.: Textural features for image classification. IEEE Trans. Syst. Man Cybern. 3, 610–621 (1973)
27. Galloway MM: Texture analysis using grey level run lengths. Comput. Graph. Image Process. 4, 172–179 (1975)
28. Chu A, et al.: Use of gray value distribution of run lengths for texture analysis. Pattern Recogn. Lett 11, 415–419 (1990)
29. Dasarathy BV, Holder EB: Image characterizations based on joint gray level run length distributions. Pattern Recogn. Lett 12, 497–502 (1991)
30. Tang X: Texture information in run-length matrices. IEEE Trans. Image Process. 7, 1602–1609 (1998) [PubMed: 18276225]
31. Amadasun M, King R: Textural features corresponding to textural properties. IEEE Trans. Syst. Man Cybern. 19, 1264–1274 (1989)
32. Hogea C, et al.: An image-driven parameter estimation problem for a reactiondiffusion glioma growth model with mass effects. J. Math. Biol 56, 793–825 (2008) [PubMed: 18026731]
33. Hogea C, et al.: A robust framework for soft tissue simulations with application to modeling brain tumor mass effect in 3D MR images. Phys. Med. Biol 52, 6893–6908 (2007) [PubMed: 18029982]
34. Hogea C, Davatzikos C, Biros G: Modeling glioma growth and mass effect in 3D MR images of the brain In: Ayache N, Ourselin S, Maeder A (eds.) MICCAI 2007. LNCS, vol. 4791, pp. 642–650. Springer, Heidelberg (2007). 10.1007/978-3-540-75757-3_78
35. Itakura H, et al.: Magnetic resonance image features identify glioblastoma phenotypic subtypes with distinct molecular pathway activities. Sci. Transl. Med 7(303), 303ra138 (2015)
36. Drabycz S, et al.: An analysis of image texture, tumor location, and MGMT promoter methylation in glioblastoma using magnetic resonance imaging. Neuroimage 49(2), 1398–1405 (2010) [PubMed: 19796694]
37. Gooya A, et al.: GLISTR: glioma image segmentation and registration. IEEE Trans. Med. Imaging 31(10), 1941–1954 (2012) [PubMed: 22907965]
38. Kendall MG: The advanced theory of statistics In: The Advanced Theory of Statistics, 2nd edn (1946)

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

39. He P, et al.: Knock-down of endogenous bornavirus-like nucleoprotein 1 inhibits cell growth and induces apoptosis in human oligodendroglia cells. Int. J. Mol. Sci 17(4), 435 (2016) [PubMed: 27023521]

40. Duhem-Tonnelle V, et al.: Differential distribution of erbB receptors in human glioblastoma multiforme: expression of erbB3 in CD133-positive putative cancer stem cells. J. Neuropathol. Exp. Neurol 69(6), 606–622 (2010) [PubMed: 20467331]

41. Carminati PO, et al.: Alterations in gene expression profiles correlated with cisplatin cytotoxicity in the glioma U343 cell line. Genet. Mol. Biol 33(1), 159–168 (2010) [PubMed: 21637621]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

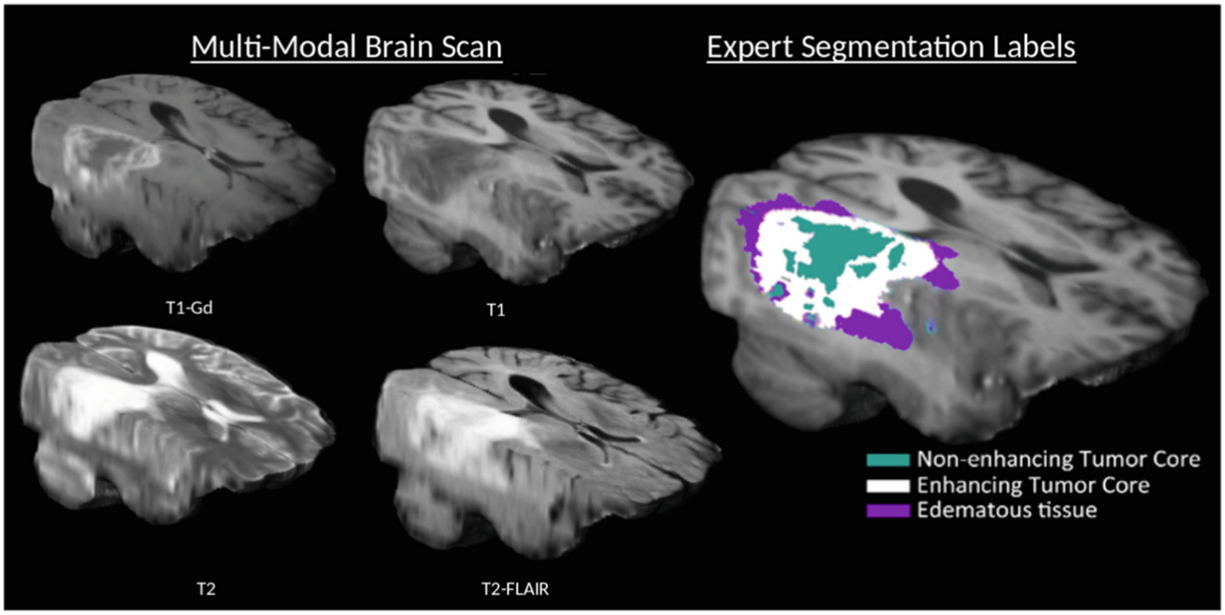

Multi-Modal Brain Scan
Expert Segmentation Labels
T1-Gd
T1
T2
T2-FLAIR
🟩 Non-enhancing Tumor Core
⬜ Enhancing Tumor Core
🟪 Edematous tissue

Fig. 1.

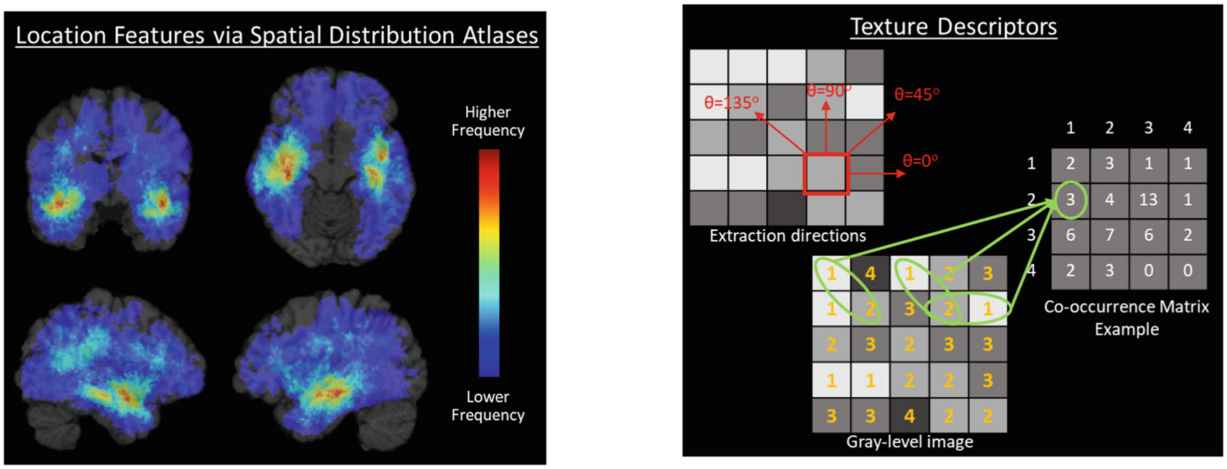
Example of a multi-modal MRI brain scan and its corresponding expert segmentation labels.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript



Location Features via Spatial Distribution Atlases
Higher Frequency
Lower Frequency
Texture Descriptors
θ=135°
θ=90°
θ=45°
θ=0°
Extraction directions
Gray-level image
Co-occurrence Matrix Example

1	4	1	2	3
1	2	3	2	1
2	3	2	3	3
1	1	2	2	3
3	3	4	2	2

	1	2	3	4
1	2	3	1	1
2	3	4	13	1
3	6	7	6	2
4	2	3	0	0

Fig. 2.

Illustrative examples of spatial distribution (left) and texture (right) patterns.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

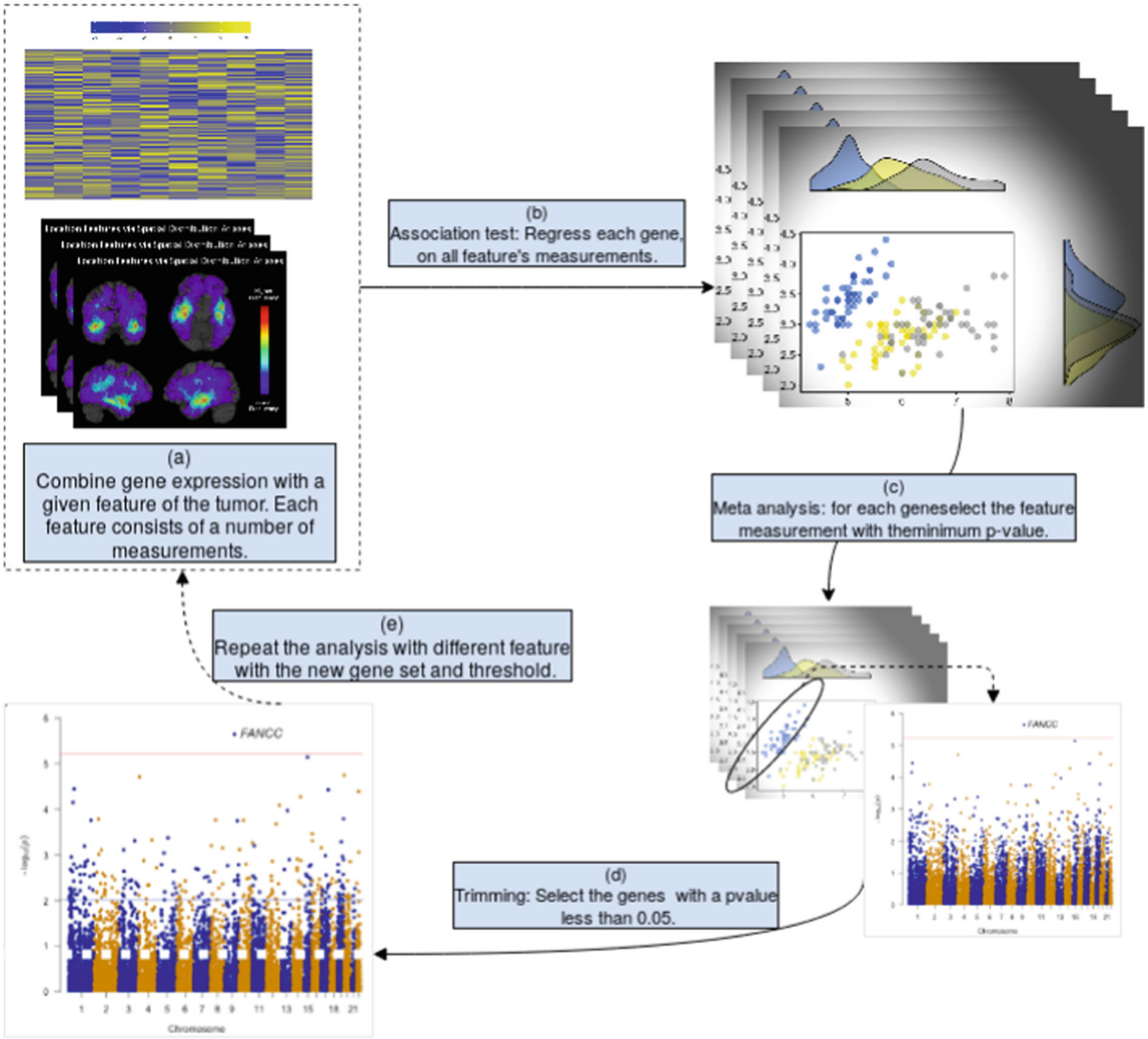

(a)
Combine gene expression with a given feature of the tumor. Each feature consists of a number of measurements.
(b)
Association test: Regress each gene, on all feature's measurements.
(c)
Meta analysis: for each geneselect the feature measurement with theminimum p-value.
(d)
Trimming: Select the genes with a pvalue less than 0.05.
(e)
Repeat the analysis with different feature with the new gene set and threshold.
FANCC
Chromosome

Fig. 3.
Schematic representation of the study's analysis workflow.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

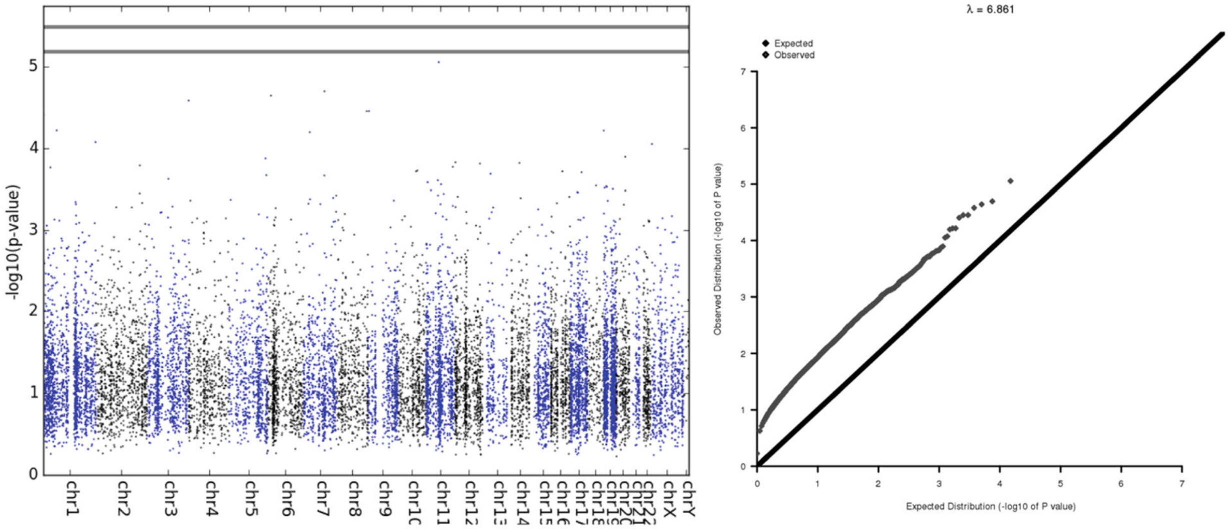


Fig. 4.

A Manhattan (left) and qq-plot (right) of the associations between the tumor spatial distribution patterns, and gene expressions. The plot is showing the metaanalysis results.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

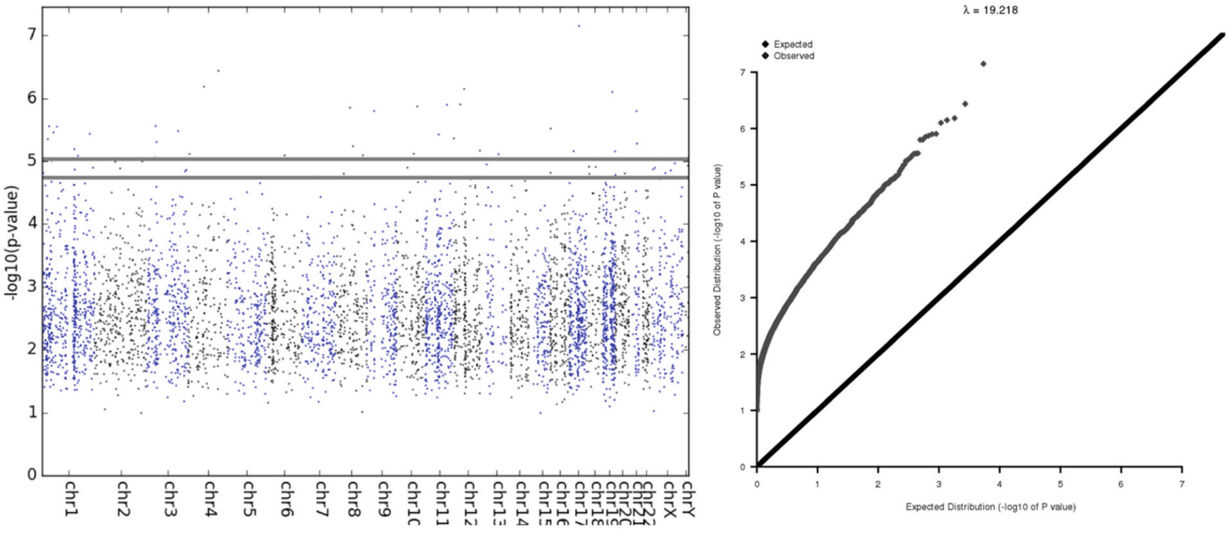


Fig. 5.

A Manhattan (left) and qq-plot (right) of the associations between the tumor texture features, and gene expressions. The plot is showing the meta-analysis results.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 1.

Number of genes, 5% and 10% thresholds used at each stage of the analysis.

Feature	No. genes used	5% threshold	10% threshold	Genes after trim
Location	15009	0.000003331 ($3.3e^{-6}$)	0.000006663 ($6.7e^{-6}$)	5401
Texture	5401	0.000009258 ($9.3e^{-6}$)	0.000018515 ($1.9e^{-5}$)	5370

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 2.

Number and percentage of patients with tumor per brain region

Loc	Vent TC	Vent ED	Frontal	Temporal	Parietal	Basal	Insula	Fornix	Occipital	Cerebellum	Brainstem
No	88	88	63	70	62	55	43	26	35	8	24
%	100.00	100.00	71.59	79.5B	70.45	62.50	48.86	29.55	39.77	9.09	27.27

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 3.

Top 10 genes: non-parametric association between genes and brain tumor location features in glioblastoma ordered according to the absolute value of r_S.

Gene	Start	End	r_s	p-value	Spatial pattern	Chromosome
TCN1	59620272	59634048	0.454	8.814e–06	DIST_Vent_TC	chr11
OR2AE1	99473609	99474680	−0.438	2.010e–05	Basal_G	chr7
KIF13A	17759413	17987854	−0.435	2.271e–05	Basal_G	chr6
NCBP2	196662272	196669468	0.432	2.619e–05	Occipital	chr3
RLN2	5299867	5304969	0.426	3.527e–05	Basal_G	chr9
KCNK9	140613080	140715299	0.426	3.533e–05	Parietal	chr8
B3GALT6	1167628	1170421	−0.423	3.938e–05	Brain_stem	chr1
FOXD3	63788729	63790797	0.414	6.0483e–05	Parietal	chr1
KISS1R	917286	921015	0.414	6.078e–05	Brain_stem	chr19
PLEKHA8	30067019	30170096	−0.413	6.362–05	Insula	chr7

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 4.

Top 10 significant genes associated with texture features of specific GBM sub-regions from specific modalities ordered according to the absolute value of r_s.

Gene	Start	End	r_s	p-value	Feature	Sub-region (MRI)	Chromosome
LRRC46	45908992	45915079	0.537	7.102e–08	GLCM Variance	ED (T2)	chr17
USP38	144106069	144144983	–0.511	3.048e–07	GLSZM SZLGE	ED (T1Gd)	chr4
EPGN	75174189	75181024	0.501	6.542e–07	GLSZM LGZE	ED (T1Gd)	chr4
TUBA1C	49582518	49667114	0.4999	7.096e–07	GLRLM RLY	NET (Tl)	chr12
ZNF284	44576296	44593766	–0.498	7.907e–07	GLRLM LGRE	NET (T1Gd)	chr19
IPO8	30781921	30848920	–0.490	1.243e–06	GLRLM GLV	ET (T2)	chr12
MMP7	102391238	102401484	0.490	1.260e–06	GLCM Auto Corr	ET (T1Gd)	chr11
TLL2	98124362	98273675	0.489	1.342e–06	GLSZM LGZE	NET (T1Gd)	chr10
TRIM55	67039130	67087720	0.488	1.408e–06	GLSZM LGZE	ED (T1Gd)	chr8
UBAP1	34179002	34252521	–0.486	1.582e–06	GLSZM SZLGE	ET (T2)	chr9

nature genetics

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.

Spatiotemporal genomic architecture informs precision oncology in glioblastoma

Jin-Ku Lee[1,2,13], Jiguang Wang[3–5,13], Jason K Sa[1,2,6,13], Erik Ladewig[3,4,13], Hae-Ock Lee[7], In-Hee Lee[1,2], Hyun Ju Kang[1,2], Daniel S Rosenbloom[3,4], Pablo G Camara[3,4], Zhaoqi Liu[3,4], Patrick van Nieuwenhuizen[3,4], Sang Won Jung[1,2,6], Seung Won Choi[1,2,6], Junhyung Kim[1,2], Andrew Chen[3,4], Kyu-Tae Kim[7], Sang Shin[1,2,6], Yun Jee Seo[1,2], Jin-Mi Oh[1,2], Yong Jae Shin[1,2,7], Chul-Kee Park[8], Doo-Sik Kong[2], Ho Jun Seol[2], Andrew Blumberg[9], Jung-Il Lee[2], Antonio Iavarone[10–12], Woong-Yang Park[6,7], Raul Rabadan[3,4] & Do-Hyun Nam[1,2,6]

Precision medicine in cancer proposes that genomic characterization of tumors can inform personalized targeted therapies[1–5]. However, this proposition is complicated by spatial and temporal heterogeneity[6–14]. Here we study genomic and expression profiles across 127 multisector or longitudinal specimens from 52 individuals with glioblastoma (GBM). Using bulk and single-cell data, we find that samples from the same tumor mass share genomic and expression signatures, whereas geographically separated, multifocal tumors and/or long-term recurrent tumors are seeded from different clones. Chemical screening of patient-derived glioma cells (PDCs) shows that therapeutic response is associated with genetic similarity, and multifocal tumors that are enriched with *PIK3CA* mutations have a heterogeneous drug-response pattern. We show that targeting truncal events is more efficacious than targeting private events in reducing the tumor burden. In summary, this work demonstrates that evolutionary inference from integrated genomic analysis in multisector biopsies can inform targeted therapeutic interventions for patients with GBM.

Most clinical trials for targeted therapy in GBM have shown limited clinical success[15]. Although recent genome-wide studies evaluating regional heterogeneity[9,11] and longitudinal GBM pairs[8,12,14,16] have suggested potential evolutionary models of the tumors, there is little understanding regarding which strategies can effectively use genomic data to inform targeted therapies. To identify such strategies, we analyzed somatic variants in 127 multiregion or longitudinal tumor specimens from 52 individuals with glioma: 42 individuals from the Samsung Medical Center (SMC), Seoul, and 10 individuals from The Cancer Genome Atlas (TCGA) GBM cohort[8,14,16] (**Supplementary Table 1**). Additionally, we analyzed the transcriptomes of 83 tumor specimens from 41 individuals (bulk) and 305 single cells from 7 samples for 3 individuals. Tumors were classified into four distinct groups according to the spatial and temporal features of tissue acquisition: tumors obtained from the same location at the same time (locally adjacent), tumors obtained from different locations at the same time (multifocal/multicentric; referred to as multiple)[17], and tumors obtained from local and distant recurrences at different times (longitudinal local and distant, respectively) (**Fig. 1a**).

We inferred clonal and subclonal alterations from cancer cell fractions in multiple sectors (Online Methods and **Supplementary Table 2**). The average mutation rate was 2.2 mutations/Mb for non-hypermutated samples, which is consistent with previous studies[2,14]. *IDH1* mutations mapping to Arg132 were clonal across all of the regions in IDH-mutant tumors[2,14] (6/6). *PIK3CA* mutations were always clonal and shared by all sectors (5/5), which is consistent with the results of our previous longitudinal analysis from tumor evolution directed graphs (TEDGs), in which we found that *PIK3CA* mutations are early events[14] (**Fig. 1b** and **Supplementary Fig. 1a**,**b**). Furthermore, *FGFR3-TACC3* fusions were highly expressed in all regions from two individuals[18]. These somatic variants, which are shared by all tumor regions, represent promising therapeutic targets[14,19], as they reflect truncal alterations in the evolutionary tree that are suspected to be present among all tumor cells. In contrast, *PTEN* alterations, including copy number deletions and mutations, were shared by 10 of 20 (50%) and 5 of 7 (71.4%) tumor sectors, respectively. Likewise, *EGFR* amplifications were observed as private events, exclusive to only one or two regions of the multisector

---

[1]Institute for Refractory Cancer Research, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea. [2]Department of Neurosurgery, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea. [3]Department of Systems Biology, Columbia University, New York, New York, USA. [4]Department of Biomedical Informatics, Columbia University, New York, New York, USA. [5]Divisions of Life Science and Biomedical Engineering, Hong Kong University of Science and Technology, Hong Kong. [6]Department of Health Science and Technology, Samsung Advanced Institute for Health Science and Technology, Sungkyunkwan University, Seoul, Republic of Korea. [7]Samsung Genome Institute, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea. [8]Department of Neurosurgery, College of Medicine, Seoul National University and Seoul National University Hospital, Seoul, Republic of Korea. [9]Department of Mathematics, University of Texas, Austin, Texas, USA. [10]Institute for Cancer Genetics, Columbia University, New York, New York, USA. [11]Department of Neurology, Columbia University, New York, New York, USA. [12]Department of Pathology, Columbia University, New York, New York, USA. [13]These authors contributed equally to this work. Correspondence should be addressed to R.R. (rr2579@cumc.columbia.edu), W.-Y.P. (woongyang@skku.edu) or D.-H.N. (nsnam@skku.edu).

Received 1 August 2016; accepted 10 February 2017; published online 6 March 2017; doi:10.1038/ng.3806

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.

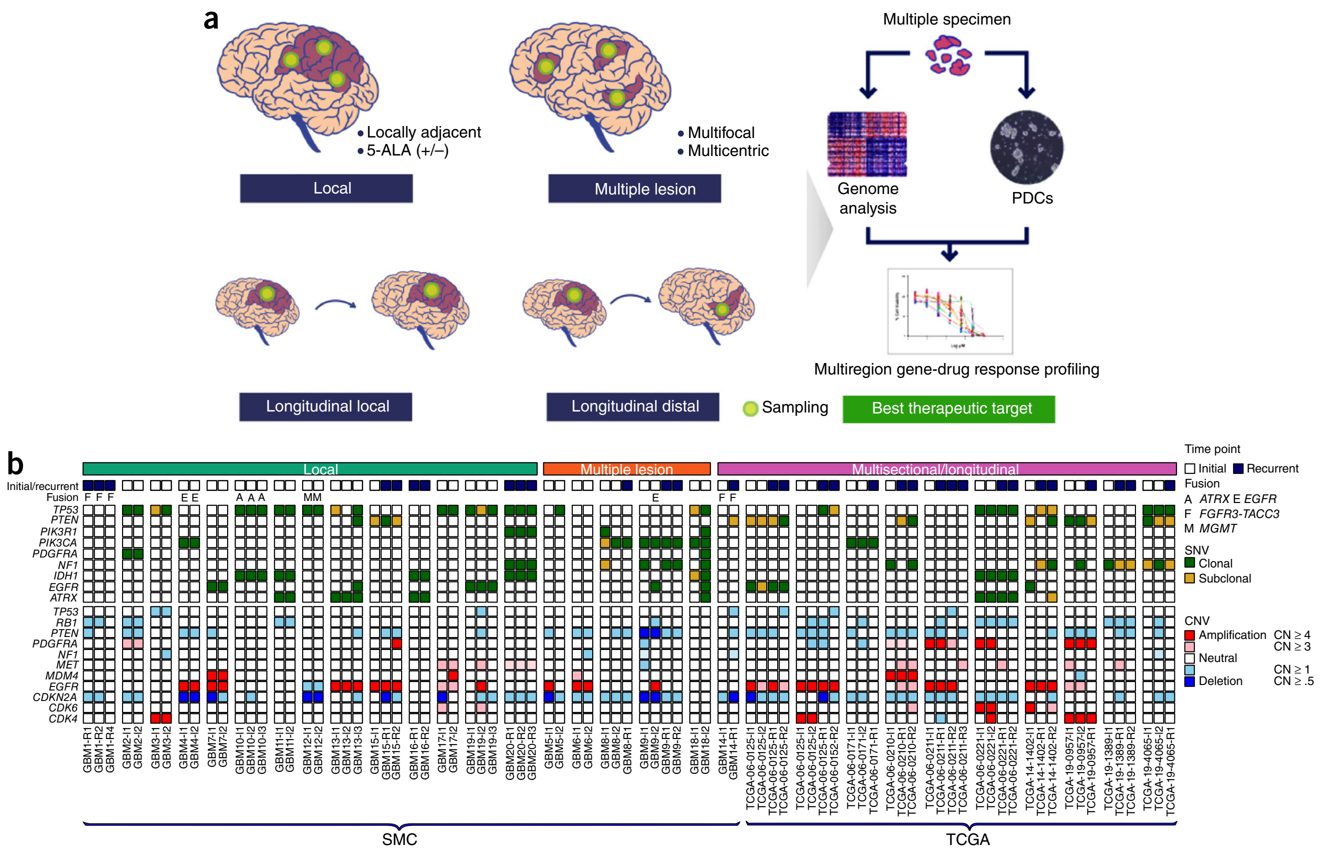


Figure 1 Mutational landscape of multiregion malignant glioma samples. (**a**) A schematic representation of glioma genomic heterogeneity and differential drug-response analysis. Human glioma specimens were acquired on the basis of their spatial order or longitudinal pairing and were subjected to genomic analysis for identification of tumor-initiating (truncal) events. PDCs, patient-derived cells; 5-ALA, 5-aminolevulinic acid. (**b**) Somatic mutations, including single-nucleotide variants (SNVs) and small insertions/deletions, copy number alterations, and gene fusions, for 83 glioma multiregion or multisector longitudinal specimens from 30 patients are shown. 34 locally adjacent tumor fragments from 14 patients, 13 multifocal/multicentric (referred to as multiple) tissues from 5 patients, and a longitudinal pair, GBM14, with leptomeningeal seeding were collected from the Samsung Medical Center (SMC). We also curated 34 multisector longitudinal tumor exomes and/or RNA sequencing data from 10 patients in The Cancer Genome Atlas (TCGA) cohort[8]. All somatic mutations called by SAVI with an allele frequency of >5% are shown. For each gene, we calculated copy number (CN) on the basis of Excavator results. Clonal alterations were determined using ABSOLUTE with a cancer cell fraction of >80% (Online Methods). CNV, copy number variation.

sample, in 4 of 15 (26.7%) of the *EGFR*-amplified tumors, including 2 multiple cases (GBM5 and GBM9). Furthermore, *EGFR* mutations were shared by 3 of 7 (42.9%) cases, including one harboring disjoint alterations, which are different genomic alterations in the same gene (GBM7-I1: p.Leu62Arg and p.Arg108Lys; GBM7-I2: p.Ala289Val and p.Cys624Ser), suggesting that partial genetic information from a single tumor biopsy can be inconclusive in assessing the potential benefit from EGFR-targeted therapy (**Supplementary Fig. 1b**,**c**).

To understand the association between spatiotemporal architecture and genetic relevance, we calculated Nei's genetic distances among multisector samples from the same individual (Online Methods). Genetic diversity was greater in multiple tumors than in locally adjacent tissues ($q = 4.7 \times 10^{-5}$, Wilcoxon rank-sum test; **Fig. 2a**), in distant recurrences than in local recurrences ($q = 1.4 \times 10^{-5}$, Wilcoxon rank-sum test; **Fig. 2a**), and in long-term recurrence than in short-term recurrences ($q = 2.9 \times 10^{-3}$, Wilcoxon rank-sum test; **Fig. 2a**). Multinomial logistic regression was applied to classify multisector sample pairs on the basis of their genomic features (**Supplementary Note**). This analysis highlighted that tumors from distant regions or long-term recurrences, separated by surgical intervals exceeding 18 months, constitute a distinct evolutionary scenario in GBM (**Fig. 2b**). In colorectal tumors, a Big Bang model[20] shows that cells from different biopsies of the same tumor share clonal and subclonal variants (**Fig. 2c**, left). In accordance with this model, samples taken from locally adjacent tumors shared a large proportion of clonal and subclonal events (**Supplementary Fig. 2a**,**b**). In contrast, multiple tumors contained fewer shared (more private) clonal mutations when compared to local tumors ($q = 1.86 \times 10^{-3}$, Wilcoxon rank-sum test; **Supplementary Fig. 2a**). We corroborated this finding by computing statistics on the space of evolutionary trees (evolutionary moduli spaces; Online Methods and **Supplementary Fig. 3**). Locally adjacent tumors clustered near the tip of the space, indicating a higher shared mutation ratio than that of multiple tumors ($P = 1.27 \times 10^{-2}$, Wilcoxon rank-sum test). These results indicate that, in contrast to locally adjacent tumors, geographically separated multifocal tumors and/or long-term recurrent tumors are seeded from distinct clones, a phenomenon we call the 'multiverse model' (**Fig. 2c**, right). Unlike the Big Bang model[20], in the multiverse model, tumor samples that are derived from different tumor masses share very few genomic alterations, indicating that tumor clones are geographically segregated at an early stage of evolution and that each clone acquires distinct private alterations, leading to the construction of multiple universes.

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.

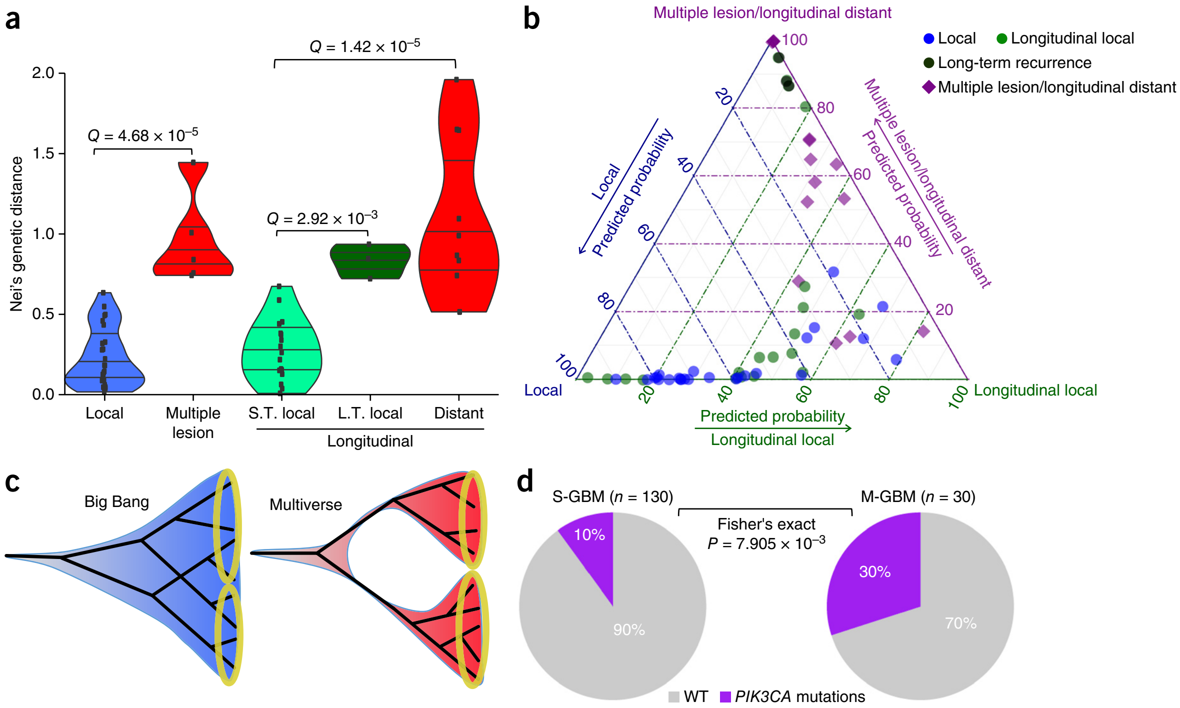


Figure 2 Comparison of genetic heterogeneity across glioma multisector and longitudinal samples. Patient samples were classified into five groups for comparative analyses: local, multiple lesion, S.T. (short-term) longitudinal local, L.T. (long-term) longitudinal local, and longitudinal distant. (**a**) Nei's genetic distances for each of the indicated groups are shown. *q* values were calculated by Wilcoxon rank-sum test and corrected for false discovery rate (FDR) using the Benjamini–Hochberg method. S.T. and L.T. local correspond to short-term (<18-month surgical interval) and long-term (≥18-month interval) recurrent tumors, respectively. The lines within each 'violin' represent the 25th, 50th, and 75th quantiles. *n* = 25 (local), 5 (multiple lesion), 17 (S.T. local), 3 (L.T. local), and 10 (distant). (**b**) Illustration of leave-one-out results from multinomial logistic regression. Each point represents one pair of samples, and the coordinates correspond to the probability that the pair is local, multiple lesion/longitudinal distant, or longitudinal local. L.T. recurrent samples were classified together with multiple lesion/longitudinal distant samples, indicating that they might follow the same evolutionary model. (**c**) Tumor evolution behind the Big Bang and multiverse models. The Big Bang model is represented as a mixture of tumor cells that share many clonal and subclonal alterations. The multiverse model is represented by a greater proportion of private events at a clonal level. (**d**) Pie charts show the frequencies of *PIK3CA* mutations in multifocal/multicentric glioblastomas (M-GBMs) (30%, 9/30) and solitary glioblastomas (S-GBMs) (10%, 13/130). WT, wild type. The *P* value was calculated using Fisher's exact test.

Next, we investigated the mutation profiles of GBMs with multifocal/multicentric lesions (M-GBMs) or solitary lesions (S-GBMs) in a total of 160 treatment-naive individuals from both the SMC and TCGA[17] cohorts (Online Methods, **Supplementary Figs. 4** and **5a**, and **Supplementary Table 3**). Notably, nonsynonymous mutations of *PIK3CA* were enriched in M-GBM (13/130 S-GBM and 9/30 M-GBM tumors; $P = 7.905 \times 10^{-3}$, Fisher's exact test; **Fig. 2d**). This conclusion remained the same in the *IDH1*-wild-type cohort (**Supplementary Fig. 5b**). *PIK3CA* induces multipotency of mammary tumors, suggesting that it has an associative role in tumor multiplicity[21]. Survival analysis indicated that both patients with M-GBM and *PIK3CA*-mutant patients had a worse prognosis than patients with S-GBM or wild-type *PIK3CA* (P = 0.0151 and 0.039, respectively, log-rank test; **Supplementary Fig. 5c**,**d**).

To further characterize the heterogeneity of expression profiles, we curated single-cell RNA–seq data from a total of seven different samples from three patients. Overall, expression-based cell subtypes were not clearly determined by location or time (**Fig. 3**), which is consistent with a previous report[10]. To make sure that this observation was not due to the limitations of this classification and to capture the transcriptional similarity among different cells, we used topological data analysis (Online Methods), a recently developed technique that summarizes and reduces the dimensionality of large data sets while retaining local high-dimensional structure[22,23].

GBM9 (**Fig. 3a–d**) consisted of samples from two initial tumors in the right and left frontal lobes and a recurrent tumor in the left frontal lobe that emerged after concurrent chemoradiotherapy (CCRT) and EGFR-targeted treatment (**Supplementary Note**). We found in bulk whole-exome sequencing (WES) and confirmed using ultra-deep sequencing (**Supplementary Table 4**) and single-cell analysis that cells from the recurrent tumor shared genomic and expression features with initial tumor cells from the left frontal lobe (**Fig. 3b** and **Supplementary Figs. 6–8**). Particularly, there were 61 somatic mutations shared by the left initial tumor and the recurrent tumor, while there were only 42 shared by the right initial tumor and the recurrent tumor. Single-cell transcriptome analysis showed *EGFR* expression predominantly in the right tumor mass, but not in the left initial and recurrent tumors (**Fig. 3c**). Different single cells harbored different *EGFR* alterations, implying that these alterations were late events during tumor evolution[21]. *PIK3CA* mutations were detected from single cells in all three samples, which is consistent with the bulk WES result that *PIK3CA* mutations are founder events (**Supplementary Figs. 1a** and **6a**). Our analysis also showed the presence of transcriptional heterogeneity within the individual samples. A subset of the left initial tumor cells was characterized by upregulation of mitotic genes that was not found in either the right or recurrent sections (**Fig. 3d**).

Additionally, we profiled *IDH1*-mutant tumor cells, which are distinguished by their 5-aminolevulinic acid (5-ALA) uptake

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.

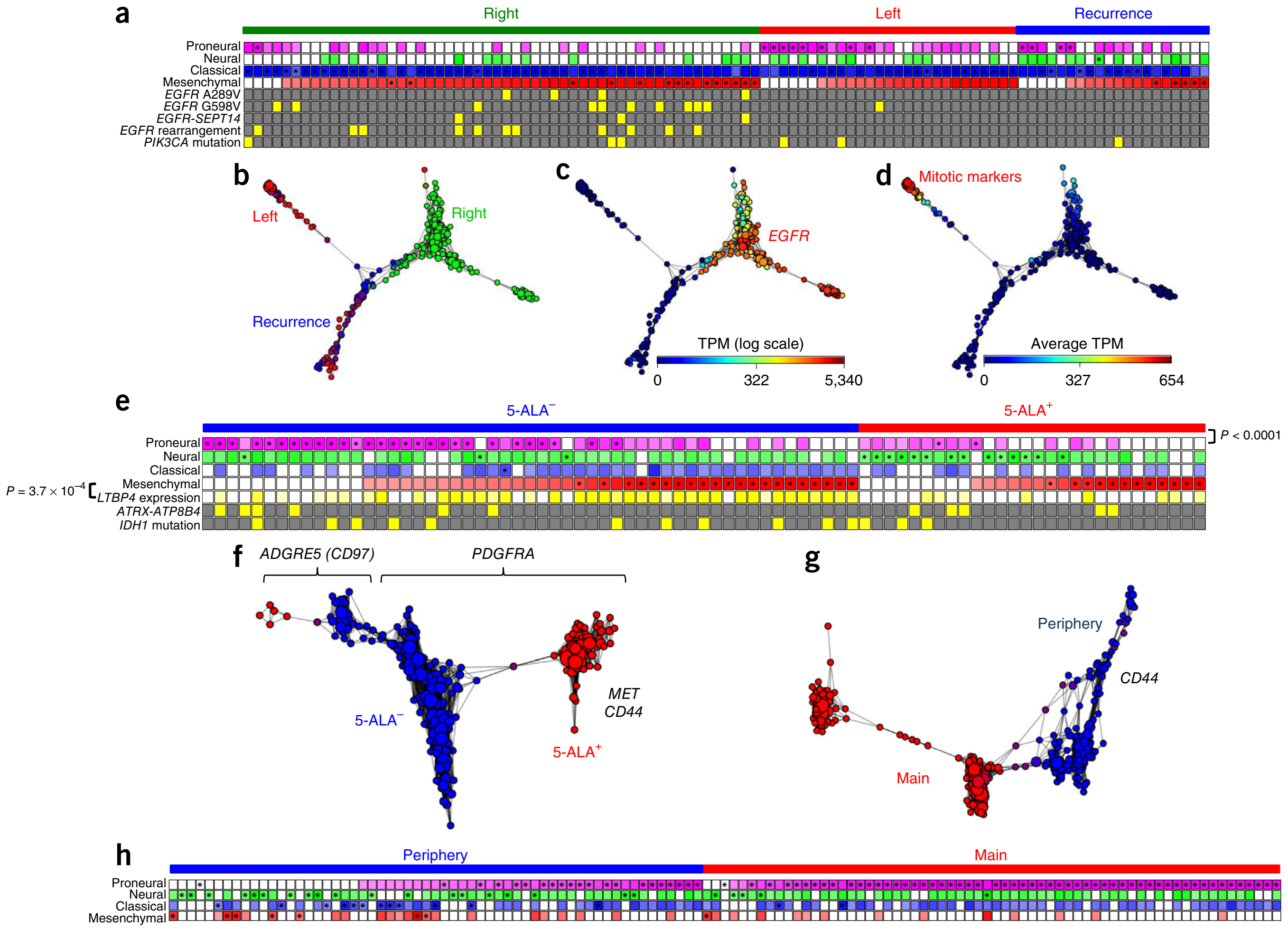


Figure 3 Single-cell transcriptomes from multiregion samples. (**a**) Expression profiles of individual tumor cells from three samples of GBM9 (left initial, right initial, and relapse) according to expression subtype[32]. For each cell, the subtype with the highest expression is marked with an asterisk. Several *EGFR* genomic alterations could be identified in the single-cell expression data (yellow), despite the abundance of missing data (gray). (**b**) Topological representation of the expression data of individual tumor cells from GBM9, labeled by sample of origin. Each node represents a set of cells with similar transcriptional profiles. A cell can appear in several nodes, and two nodes are connected by an edge if they have at least one cell in common. (**c**,**d**) Topological representations of GBM9 where tumor cells are labeled by expression of *EGFR* (**c**) and mitotic genes (**d**). TPM, transcripts per million. (**e**) Expression profiles of individual tumor cells from GBM10 (two samples: 5-ALA^{+} and 5-ALA^{-}). The *P* value for comparison of the proneural and 5-ALA profiles was obtained using Fisher's exact test. The *P* value for comparison of the mesenchymal profile and *LTBP4* expression was calculated using Spearman's correlation. *ATRX* fusion was validated by RT–PCR assays (**Supplementary Fig. 7b,c**). (**f**,**g**) Topological representations of expression data for individual tumor cells from patients GBM10 (**f**) and GBM2 (**g**), where tumors are labeled by the sample of origin. (**h**) Expression profiles of individual tumor cells from patient GBM2 shown according to GBM expression subtype. Asterisks indicate the representative subtype identities of the corresponding cells.

pattern (populations stained for tumor cellularity[24]) (GBM10; **Fig. 3e**,**f** and **Supplementary Figs. 6** and **7**). Previous glioma studies suggested that a low pathologic grade is associated with a low rate of 5-ALA uptake[24,25]; however, genomic determinants for 5-ALA uptake remain elusive. Using single-cell transcriptome analysis, we found predominant enrichment of proneural cells in the 5-ALA^{-} sample, supporting previous observations that GBM cells may evolve from proneural precursors ($P < 0.01$, Fisher's exact test; **Fig. 3e**)[26]. Also, we found enrichment of expression for several cell proliferation and migration markers in the 5-ALA^{+} section, including *MET* and *CD44* (refs. 27–29). Notably, 5-ALA^{-} tumors, which are considered to be less aggressive, are actually fully fledged tumors that harbor driver mutations and express markers for aggressive tumors (**Fig. 3e**,**f** and **Supplementary Figs. 6b**, **7**, **9a**, and **10**). Finally, we studied main tumor and resection margin samples from GBM2—a locally adjacent hypermutated case—and found distinct subpopulations of cells expressing mitotic cell markers and migration-associated genes, including *CD44* (**Fig. 3g** and **Supplementary Figs. 6**, **7**, **9**, and **10**)[28].

To investigate the influence of genetic heterogeneity on drug response, we isolated 28 PDCs from 11 individuals and screened 40 different cancer-related compounds (**Supplementary Table 5**)[4,30]. We found that Nei's genetic distance was associated with drug-response correlation ($P = 0.02$, Wilcoxon rank-sum test; **Fig. 4a**). Consistently, both distant and longitudinal samples showed significantly broader drug responses than local samples (**Fig. 4b**). We found that the PDCs from M-GBMs were more sensitive to PI3K–AKT–mTOR (PAM) pathway inhibitors than PDCs from solitary tumors were ($P = 1.872 \times 10^{-6}$, Wilcoxon rank-sum test; **Fig. 4c** and **Supplementary Fig. 11**). This indicates that PAM inhibitors could provide a clinical benefit for patients with M-GBM. In addition, we observed that PDCs from recurrent GBMs were more resistant to EGFR inhibitors than

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.

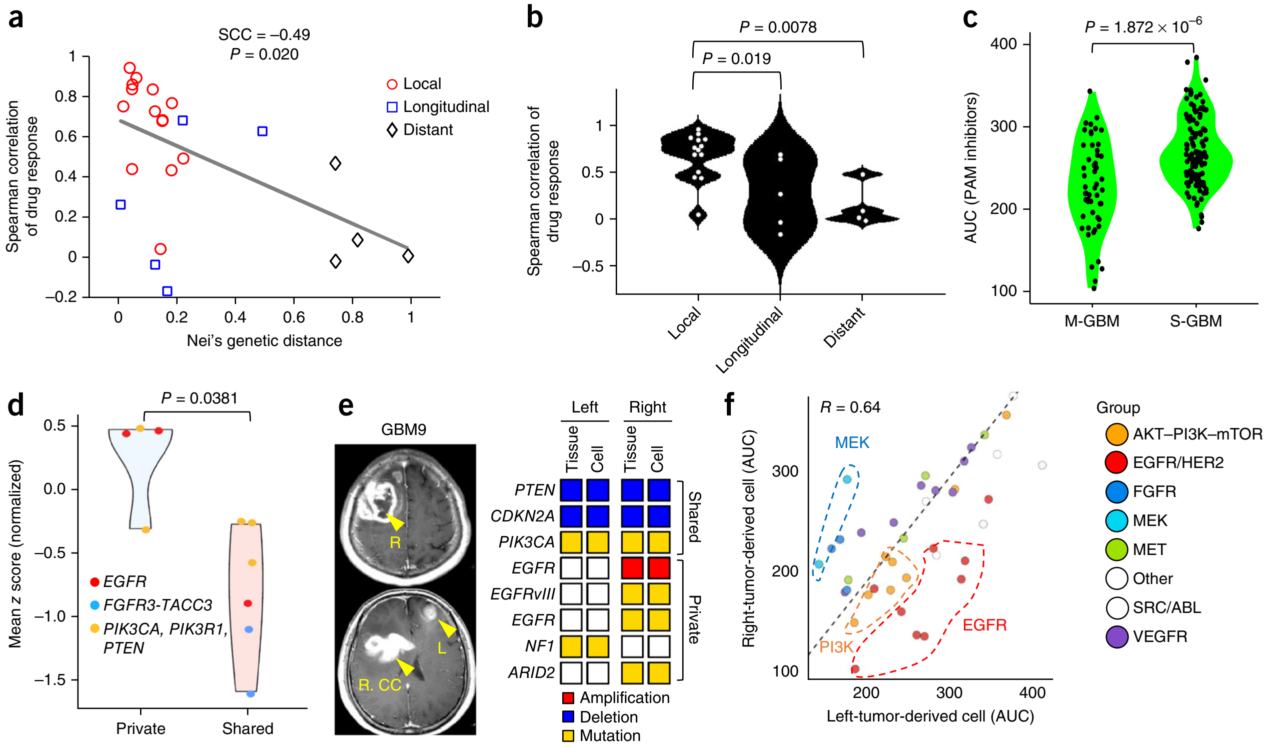


Figure 4 Chemical screening of multiregion PDCs. (**a**) PDCs were treated with 40 chemical agents targeting oncogenic signaling pathways in dilution series from 20 μM to 4.88 nM. The *x* axis shows Nei's genetic distances between fragments from the same patient, and the *y* axis shows the Spearman's correlation coefficient (SCC) of corresponding fragments based on drug sensitivities as measured by area under the curve (AUC). (**b**) Violin plots for the SCC of drug responses for the groups described in **a**. Each white dot represents a tumor sample. $n = 14$ (local), 5 (longitudinal), and 4 (distant). (**c**) Mean values of the AUCs for six PI3K–AKT–mTOR (PAM) inhibitors (BEZ235, BKM120, BYL719, AZD5363, AZD2014, and everolimus) for PDCs isolated from M-GBMs ($n = 9$) or S-GBMs ($n = 22$). Each black dot represents a mean. (**d**) The normalized *z* score for each PDC obtained from the AUCs for the indicated drug classes used in **a** and **b** was plotted when the corresponding tissues harbored associated genetic alterations, designated as 'shared' or 'private'. A genetic alteration was determined to be private when the drug-response-associated genetic alteration was private, and vice versa for the shared group. For example, *PTEN* mutations for P13K–AKT–mTOR pathway inhibitors were private, meaning they were exclusive to only one or two regions of the multisector sample. (**e**) Preoperative T1-weighted contrast-enhanced magnetic resonance images (MRIs) and key genomic alterations found in the corresponding tumors and their derivative cells for a patient with multicentric disease (GBM9). "R" indicates right-side tumors that encompassed the right frontal lobe and corpus callosum (CC). "L" indicates the left frontal lobe tumor. The preoperative MRIs show a multifocal infiltrative lesion in both the frontal lobes and CC. (**f**) Scatterplot of the AUCs for 40 cancer-targeting compounds on GBM9 PDCs that were derived from the left- and right-side tumors in the frontal lobes. Each dot represents the AUC of a given drug. The *R* value was obtained as Pearson's correlation coefficient. All *P* values in this figure were obtained using Wilcoxon rank-sum tests.

the initial tumor cells were ($P = 2.9 \times 10^{-4}$, Wilcoxon rank-sum test; **Supplementary Figs. 12** and **13**).

We hypothesized that clonal alterations found in all multisector samples (truncal alterations) represent better molecular targets than those found in only a subset of multisector samples (private alterations). In agreement with this truncal-target hypothesis (**Supplementary Video 1**), multisector PDCs were more sensitive to drugs that target shared alterations than to drugs that target private alterations ($P = 0.0381$, Wilcoxon rank-sum test; Online Methods, **Fig. 4d** and **Supplementary Figs. 14–16**). The multiverse model implies that the extensive genetic diversity of multiple tumors presents a special challenge. Accordingly, GBM9 showed a divergent genetic profile and a highly heterogeneous drug response (**Fig. 4e**,**f**). PDCs from the right-side tumor were highly sensitive to EGFR inhibitors, but not to MEK inhibitors, and vice versa for the left-side tumor. However, inhibitors of the PAM pathway were ubiquitously effective, which is consistent with our hypothesis that targeting the PAM signaling pathway could be a potent option to treat M-GBMs (**Fig. 4c**,**f**). Yet not all truncal alterations can serve as drug targets. For example, gatekeeper genes, which are necessary for tumor initiation but are no longer required for tumor maintenance, are not good candidates[31]. Although targeting subclonal mutations shows a limited effect, patients might still benefit from the elimination of a subclone that has a bystander effect on surrounding cells.

In conclusion, on the basis of comprehensive bulk and single-cell analyses, we have proposed a multiverse model to interpret the evolution of multiple GBMs. We showed that M-GBMs are more genetically diverse than locally adjacent tumors and that genetic similarity between multiregion samples is associated with consistent drug response. Specifically, we found an enrichment of *PIK3CA* mutations in M-GBMs, and we found that inhibitors of the PAM pathway are more effective in PDCs from this cohort. These findings support the truncal-target hypothesis, which states that truncal mutations can inform more effective therapies.

URLs. R nnet package, multinom function, https://CRAN.R-project.org/package=nnet; Mapper algorithm, http://danifold.net/mapper; http://github.com/MLWave/kepler-mapper.

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

This work was supported by a grant of the Korea Health Technology R&D project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (HI14C3418). R.R. acknowledges funding from the NIH (U54 CA193313, R01 CA185486, R01 CA179044).
J.W. is supported by Precision Medicine Fellowship (UL1 TR000040). E.L. is supported by NIH (F99 CA212478) and Cancer Biology Training Program (T32 CA09503).

AUTHOR CONTRIBUTIONS

J.-K.L., J.W., J.K.S., and E.L. are co-first authors. J.-K.L., J.W., J.K.S., and E.L. performed the majority of experiments and analyses. W.-Y.P., H.-O.L., K.-T.K., P.G.C., and P.v.N. performed experiments and analyses for the single-cell transcriptome. D.S.R., Z.L., A.B., A.C., Y.J.S., and S.S. conducted several experiments and analyses. D.-H.N., D.-S.K., H.J.S., C.-K.P., and J.-I.L. provided surgical specimens. S.W.J., S.W.C., and J.K. helped interpret clinical data. W.-Y.P., I.-H.L., Y. J.S., J.-M.O., and H.J.K. organized and processed specimens and genome data. J.-K.L., J.W., J.K.S., E.L., and P.G.C. wrote the manuscript with feedback from R.R., A.I., and D.-H.N. D.-H.N. and R.R. designed and supervised the entire project.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Hamburg, M.A. & Collins, F.S. The path to personalized medicine. *N. Engl. J. Med.* **363**, 301–304 (2010).
2. Brennan, C.W. *et al.* The somatic genomic landscape of glioblastoma. *Cell* **155**, 462–477 (2013).
3. Ceccarelli, M. *et al.* Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma. *Cell* **164**, 550–563 (2016).
4. Cloughesy, T.F., Cavenee, W.K. & Mischel, P.S. Glioblastoma: from molecular pathology to targeted treatment. *Annu. Rev. Pathol.* **9**, 1–25 (2014).
5. Frattini, V. *et al.* The integrated landscape of driver genomic alterations in glioblastoma. *Nat. Genet.* **45**, 1141–1149 (2013).
6. Snuderl, M. *et al.* Mosaic amplification of multiple receptor tyrosine kinase genes in glioblastoma. *Cancer Cell* **20**, 810–817 (2011).
7. Szerlip, N.J. *et al.* Intratumoral heterogeneity of receptor tyrosine kinases *EGFR* and *PDGFRA* amplification in glioblastoma defines subpopulations with distinct growth factor response. *Proc. Natl. Acad. Sci. USA* **109**, 3041–3046 (2012).
8. Kim, H. *et al.* Whole-genome and multisector exome sequencing of primary and post-treatment glioblastoma reveals patterns of tumor evolution. *Genome Res.* **25**, 316–327 (2015).
9. Sottoriva, A. *et al.* Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. *Proc. Natl. Acad. Sci. USA* **110**, 4009–4014 (2013).
10. Patel, A.P. *et al.* Single-cell RNA–seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**, 1396–1401 (2014).
11. Kumar, A. *et al.* Deep sequencing of multiple regions of glial tumors reveals spatial heterogeneity for mutations in clinically relevant genes. *Genome Biol.* **15**, 530 (2014).
12. Johnson, B.E. *et al.* Mutational analysis reveals the origin and therapy-driven evolution of recurrent glioma. *Science* **343**, 189–193 (2014).
13. Yates, L.R. *et al.* Subclonal diversification of primary breast cancer revealed by multiregion sequencing. *Nat. Med.* **21**, 751–759 (2015).
14. Wang, J. *et al.* Clonal evolution of glioblastoma under therapy. *Nat. Genet.* **48**, 768–776 (2016).
15. Ohka, F., Natsume, A. & Wakabayashi, T. Current trends in targeted therapies for glioblastoma multiforme. *Neurol. Res. Int.* **2012**, 878425 (2012).
16. Kim, J. *et al.* Spatiotemporal evolution of the primary glioblastoma genome. *Cancer Cell* **28**, 318–328 (2015).
17. Liu, Q. *et al.* Genetic, epigenetic, and molecular landscapes of multifocal and multicentric glioblastoma. *Acta Neuropathol.* **130**, 587–597 (2015).
18. Singh, D. *et al.* Transforming fusions of *FGFR* and *TACC* genes in human glioblastoma. *Science* **337**, 1231–1235 (2012).
19. Hartmann, C., Bartels, G., Gehlhaar, C., Holtkamp, N. & von Deimling, A. *PIK3CA* mutations in glioblastoma multiforme. *Acta Neuropathol.* **109**, 639–642 (2005).
20. Sottoriva, A. *et al.* A Big Bang model of human colorectal tumor growth. *Nat. Genet.* **47**, 209–216 (2015).
21. Koren, S. *et al.* *PIK3CA*$_{\text{H1047R}}$ induces multipotency and multi-lineage mammary tumours. *Nature* **525**, 114–118 (2015).
22. Carlsson, G. Topology and data. *Bull. Amer. Math. Soc.* **46**, 255–308 (2009).
23. Singh, G., Mémoli, F. & Carlsson, G.E. Topological methods for the analysis of high dimensional data sets and 3D object recognition. in *SPBG* (eds. Botsch, M. & Pajarola, R.) 91–100 (Citeseer, 2007).
24. Moiyadi, A., Syed, P. & Srivastava, S. Fluorescence-guided surgery of malignant gliomas based on 5-aminolevulinic acid: paradigm shifts but not a panacea. *Nat. Rev. Cancer* **14**, 146 (2014).
25. Stummer, W. *et al.* Fluorescence-guided surgery with 5-aminolevulinic acid for resection of malignant glioma: a randomised controlled multicentre phase III trial. *Lancet Oncol.* **7**, 392–401 (2006).
26. Ozawa, T. *et al.* Most human non-GCIMP glioblastoma subtypes evolve from a common proneural-like precursor glioma. *Cancer Cell* **26**, 288–300 (2014).
27. Merzak, A., Koocheckpour, S. & Pilkington, G.J. CD44 mediates human glioma cell adhesion and invasion *in vitro*. *Cancer Res.* **54**, 3988–3992 (1994).
28. Yoshida, T., Matsuda, Y., Naito, Z. & Ishiwata, T. CD44 in human glioma correlates with histopathological grade and cell migration. *Pathol. Int.* **62**, 463–470 (2012).
29. Joo, K.M. *et al.* MET signaling regulates glioblastoma stem cells. *Cancer Res.* **72**, 3828–3838 (2012).
30. Lemmon, M.A. & Schlessinger, J. Cell signaling by receptor tyrosine kinases. *Cell* **141**, 1117–1134 (2010).
31. Kinzler, K.W. & Vogelstein, B. Lessons from hereditary colorectal cancer. *Cell* **87**, 159–170 (1996).
32. Verhaak, R.G. *et al.* Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in *PDGFRA*, *IDH1*, *EGFR*, and *NF1*. *Cancer Cell* **17**, 98–110 (2010).

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.

ONLINE METHODS

Glioma specimens and their derivative cells. After receiving informed consents, glioma specimens and clinical records were obtained from patients undergoing surgery at Samsung Medical Center (SMC) or Seoul National University Hospital (SNUH) in accordance with its institutional review board (IRB file no. 2010-04-004). Surgical samples measuring approximately 5 × 5 × 5 mm^3 were snap-frozen using liquid nitrogen for genomic analysis. We also curated whole-exome and/or RNA sequencing of 33 multisector specimens from 10 glioblastoma (GBM) patients in The Cancer Genome Atlas (TCGA) cohort[8] and 22 previously reported GBM longitudinal pairs[14,16]. To investigate the genomic characteristics of solitary and multifocal/multicentric GBMs, we curated exome sequencing data for 83 and 77 tumors with matched normal DNA from the SMC and TCGA cohorts[8,17], respectively. Portions of the surgical samples were enzymatically dissociated into single cells, following the procedures reported previously with modification of immune cell depletion[33]. Tumor cells were cultured in neurobasal medium with N2 and B27 supplements (0.5× each, half of the suggested working concentration; Invitrogen), human recombinant basic fibroblast growth factor (bFGF), and epidermal growth factor (EGF; 20 ng/ml each; R&D Systems). The patient-derived cells (PDCs) used here had shown no obvious contamination of mycoplasma.

Radiological evaluation. Both T1-weighted contrast enhancement (T1CE) and fluid-attenuated inversion recovery (FLAIR)/T2 axial images of 160 treatment-naive GBMs (83 and 77 tumors from the SMC and TCGA cohorts, respectively) were reviewed. Magnetic resonance images (MRIs) of tumors from the TCGA cohort have been obtained from The Cancer Imaging Archive (TCIA) website[17,34,35]. We excluded cases with any evidence of prior neurosurgical intervention except biopsy, lack of treatment history, or loss of T1CE or FLAIR/T2 images. To distinguish the multifocal/multicentric GBMs (M-GBMs) from solitary ones (S-GBMs), we adapted annotations from the VASARI feature set for human glioma[36]. According to the VASARI feature set, m-GBMs are defined as having at least one region of tumor, either enhancing or nonenhancing, that is not contiguous with the main lesion and is outside of the region of signal abnormality (edema) surrounding the main mass[17,37,38]. When a FLAIR/T2 high-signal-intensity lesion resides outside of the T1CR lesion, it is considered a separate tumor focus and is counted as a multifocal tumor in our study[39,40]. In contrast, tumors that present separate contrast-enhancement lesions within the FLAIR/T2 high-signal-intensity background are considered as solitary ones.

Whole-exome sequencing. An Agilent SureSelect kit was used to capture the exonic DNA fragments. An Illumina HiSeq 2000 instrument was used for sequencing and generated 2 × 101-bp paired-end reads.

Somatic mutation. The sequenced reads in the FASTQ files were aligned to the human genome assembly (hg19) using Burrows–Wheeler aligner version 0.6.2. The initial alignment BAM files were subjected to conventional preprocessing before mutation calling: sorting, removing duplicated reads, locally realigning reads around potential small indels, and recalibrating base quality scores using SAMtools, Picard version 1.73, and Genome Analysis Toolkit (GATK) version 2.5.2. We used MuTect (version 1.1.4) and Somatic IndelDetector (GATK version 2.2) to make high-confidence predictions on somatic mutations from the neoplastic and non-neoplastic tissue pairs. Variant Effect Predictor (VEP) version 73 was used to annotate the called somatic mutations. Additionally, we ran SAVI (Statistical Variant Identification) software to call somatic variants and indels in order to for refine the mutation calls from the above pipeline.

Copy number. Excavator was used to generate estimated copy number alterations in a tumor specimen in comparison with its matching, non-neoplastic part. For each gene, we calculated copy number = 2^{x+1}, where x is the segmentation mean from Excavator, which is defined as the $\log_2$ (fold change) in the tumor divided by the normal sample. The gene was labeled as 'amplified' when the copy number was 3 and 'deleted' when it was ≤1.

Cancer cell fractions and clonality. We ran ABSOLUTE[41] using input of genomic variants and copy number data to infer sample purity and cancer cell fractions (CCFs) and removed those that had <20% purity. We considered mutations as clonal if they were indicated as clonal in ABSOLUTE and had a CCF of at least 80% or if they had a CCF of 100% and were not marked as clonal or subclonal. The ABSOLUTE CCF estimates with regard to hypermutated samples appeared disproportionately subclonal in sample GBM18 initial and in TCGA-14-1402 second recurrence; we reasoned that the large mutational load might skew estimates. In hypermutated samples, treatment-associated mutation coupled with defects in mismatch repair is deemed largely responsible for a majority of observed mutations. Therefore, mutations having CCFs greater than or equal to the maximum mismatch repair CCF were marked clonal in these two samples.

If a mutation was found to be clonal in all sectors of a patient's tumor, it was inferred to be clonal throughout the entire tumor. We investigated the number of sequenced tumor sectors or cores needed to obtain a reasonable false discovery rate (FDR) for this inference of clonality. We analyzed glioma patient LGG174, published recently by Suzuki *et al.*, where nine sectors from different locations of the same tumor mass were sequenced[42]. On the basis of Figure 7b from that paper[42], 13 mutations have a high tumor cell fraction (>60%) shared by all samples. To relate the number of sectors that were sequenced to the number of mutations deemed to be clonal tumor-wide, we exhausted all possible subsampling strategies (number of cores $k = 1, 2, \ldots, 9$) and calculated the reported clonal mutations based on k cores. For example, if there are two cores ($k = 2$), there are $C_2^9 = 36$ potential sampling strategies. We found that 22 of 36 sampling strategies contained no false discoveries in identifying clonal mutation. For each value of k, we calculated the FDR (**Supplementary Fig. 17**). Almost 90% of clonal mutations identified by two-core sequencing are true clonal mutations, and over 95% identified by three-core sequencing are true clonal mutations.

Nei's genetic distances. Nei's genetic distance is used in population genetics to assess the similarity between populations, taking into account heterogeneity within populations. Samples containing the same spatial or longitudinal category (local, 5-ALA, multiple lesion, longitudinal local, longitudinal distant) were retained for statistical comparisons. We calculated Nei's genetic distance of CCF for each patient's sample as follows. Let x be all CCFs of sample 1 and y be all CCFs of sample 2.

$$D = -\log \frac{\sum(x_i y_i + (1 - x_i)(1 - y_i))}{\sqrt{(\sum x_i^2 + (1 - x_i)^2)(\sum y_i^2 + 1 - y_i)^2)}}$$

The multiverse model of tumor evolution. We found an increased Nei's genetic distance in multifocal/multicentric biopsies when compared with those that were locally adjacent. In addition, private clonal mutations appear frequently in multisectional and distant longitudinal samples, but are infrequent in local samples (**Supplementary Fig. 2a**). This spurred a hypothesis that specific early event(s) can give rise to distinct mutational profiles in spatially separated tumors (**Fig. 2a**). These differences in mutational load suggested that distinct tumor profiles might arise in separate 'universes' of clones rather than from one large growth period followed by diversification.

For each somatic mutation, we recorded the clonal status as determined by ABSOLUTE and whether the mutation is shared or private, or the clonal status changed between biopsies. Mutations are then classified into five patterns with respect to every available pair of a patient's samples. The mutational classes were labeled as the following: CC (clonal–clonal), CS (clonal–subclonal), SS (subclonal–subclonal), CX (clonal–absent), or SX (subclonal–absent). The order of the sample pair was not important: a mutation that was clonal in one sample and subclonal in the other was marked "CS," regardless of sample identity.

These mutational classifications were used to predict whether the spatiotemporal configuration of a sample pair fell into one of three groups: locally adjacent, local longitudinal, or multisectional/distant longitudinal. The fractions of mutations in a sample pair that fit each of the five patterns were used as features in a multinomial logistic regression. Predictions were then made using leave-one-out cross-validation.

Mutational pairs plotted on the simplex allowed us to visually separate multisectional/distant longitudinal, locally adjacent, or local longitudinal sections in agreement with most of our MRI classifications. The simplex

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.

axes represent the predicted probabilities of outcomes for each observation. The sample layout contained three local longitudinal outliers closest to the multisectional point of the simplex. The time interval between surgeries for the three pairs was 18 months or more. Moreover, their Nei distances were significantly different from all other sections ($P = 0.01652$). We labeled all samples exceeding surgical intervals of 18 months as long-term recurrence and colored them in dark green. Analysis was performed in the R computing environment using the multinom function from the nnet package (see "Data availability").

Isolation of single cells and RNA sequencing. We adopted the C1TM Single-Cell Auto Prep System (Fluidigm) with the SMARTer kit (Clontech) to generate cDNAs from single cells. Cells were captured as a single isolate on a C1 chip (17–25 μm), as determined from bright-field images obtained under 100× magnification using an Axiovert200 inverted microscope (Carl Zeiss). RNAs from pooled samples were also processed using the SMARTer kit with 10 ng of starting materials. Libraries were generated using the Nextera XT DNA Sample Prep Kit (Illumina) and sequenced on the HiSeq 2500 using the 100-bp paired-end mode of the TruSeq Rapid PE Cluster kit and TruSeq Rapid SBS kit. Before mapping RNA sequencing reads to the reference, reads were filtered out at Q33 by using Trimmomatic-0.30. Transcripts per million (TPM) values were calculated from each single cell (as if they were different samples) using RSEM (version 1.2.25)[43] and are expressed as $\log_2$ (1 + TPM).

Gene fusion detection. Chimerascan was applied to generate a list of candidate gene fusions[44]. For bulk sequencing, only previously reported in-frame, highly expressed fusions, such as *FGFR3-TACC3* (ref. 18), *MGMT* fusion[14], *EGFR-SEPT14* (ref. 5) and *ATRX* fusion, were considered in this manuscript. For single-cell fusion analysis, if a fusion was highly expressed and independently detected in multiple cells, the fusion was reported.

Expression-based subtype determination. Gene expression was measured by RSEM and then $\log_2$ transformed. To determine the expression-based subtype of GBM cells, we first calculated z scores for gene expression data across samples and then applied ssGSEA (version gsea2-2.2.1) to the normalized expression profile. For each cell, we ranked all genes on the basis of their expression values to create a .rnk file as input for the software GseaPreranked. An enrichment score was computed for all four subtypes initially defined in Verhaak *et al.*[32]. The subtype with the maximal enrichment score was used as the representative subtype for each cell.

Topological data analysis using single-cell transcriptomes. We filtered out normal cells on the basis of their expression profile. To that end, we considered the expression signatures of normal oligodendrocytes, neurons, astrocytes[45], microglia[46], endothelial cells[47], T cells[48], and other immune cells[48,49], and we used a Gaussian mixture model to classify individual cells according to their expression profile. 94/133, 82/85, and 90/137 cells for GBM9, GBM10, and GBM2, respectively, were classified as tumor cells. After normalizing the gene expression level by dividing by the total number of reads in each cell to eliminate potential bias caused by batch effect, we built topological representations of these single-cell data using the Mapper algorithm[23], as implemented by Ayasdi. Open-source implementations of this algorithm are also available (see "Data availability"). We used the first two components of multidimensional scaling (MDS) as auxiliary functions for the algorithm. The output of Mapper is a low-dimensional network representation of the data, where nodes represent sets of cells with similar global transcriptional profiles (as measured by correlation of the expression levels of the 2,000 genes with the highest variance across each patient). We identified individual genes that had an expression pattern localized in the network and used these to determine the subclonal structure of the samples at the level of expression.

PDC-based chemical screening and analysis. PDCs grown in serum-free medium were seeded in 384-well plates at a density of 500 cells per well in duplicate or triplicate for each treatment. The drug panel consisted of 40 anticancer agents (Selleckchem) targeting oncogenic signals. Two hours after plating, PDCs were treated with drugs in a fourfold and seven-point serial dilution series from 20 μM to 4.88 nM using the Janus Automated Workstation (PerkinElmer). After 6 d of incubation at 37 °C in a 5% CO_2 humidified incubator, cell viability was analyzed using an ATP-monitoring system based on firefly luciferase (ATPLit 1step, PerkinElmer). The number of viable cells was estimated using the EnVision Multilabel Reader (PerkinElmer). DMSO was also included as a control in each plate. Controls were used for the calculation of relative cell viability for each plate, and normalization was performed on a per-plate basis. Dose-response curve (DRC) fitting was performed using GraphPad Prism 5 (GraphPad) and evaluated by measuring the area under the curve (AUC) of the DRC. After normalization, best-fit lines were determined and the AUC value of each curve was calculated using GraphPad Prism, ignoring regions defined by fewer than two peaks.

Cell viability was determined via calculating AUC values of DRCs with exclusion of nonconvergent fits[50].

Moduli space analysis. To illustrate the evolution histories of tumors in GBM patients, we applied moduli space analysis[51] in local and multiple groups of patients. Multiregion pairs were compared to calculate the number of shared and private mutations. In this analysis, clonal mutations were separated on the basis of their allele frequencies. Sector pairs were put in the left sphere on the basis of the number of shared and private mutations with high-allele-frequency mutations (>20%), whereas the same number of pairs were put in the right sphere on the basis of mutations with low-allele-frequency mutations (<20%). The same analysis was repeated using CCF instead of mutational allele fraction.

Immunohistochemistry. Tissue specimens were fixed by formalin and embedded in paraffin. Paraffin-embedded sections were treated with 0.3% hydrogen peroxide to block endogenous peroxidase activity, and antigens were retrieved by heating sections in 10 mM sodium citrate (pH 6.0) at 95 °C for 30 min. Sections were incubated with primary antibodies overnight at 4 °C, biotinylated secondary antibodies for 1 h at room temperature, and avidin–biotin complex for 1 h at room temperature.

Protein blotting. GBM PDCs were washed with cold PBS and harvested in lysis buffer (150 mM sodium chloride, 1% Triton X-100, 1% sodium deoxycholate, 0.1% SDS, 50 mM Tris-HCl, and 2 mM EDTA), and a protease and phosphatase inhibitor cocktail was added (Thermo Scientific). Insoluble materials were removed by centrifugation at 13,500*g* for 15 min at 4 °C. Proteins were separated by SDS–PAGE. Immunoblotting was performed using antibodies against indicated proteins.

Limiting dilution assays. GBM PDCs were dissociated into single-cell suspensions and then plated into 96-well plates at 1–250 cells per well. Cells were incubated at 37 °C for 1–2 weeks. At the time of quantification, each well was examined for formation of neurosphere-like cell aggregates. Statistical significance was evaluated using extreme limiting dilution analysis (ELDA; Walter+Eliza Hall Bioinformatics).

Gene fusion validation. Validation of gene fusion transcripts were performed by RT–PCR assays. Total RNA was extracted from tissues by AllPrep DNA/RNA Mini Kit according to the manufacturer's instructions (Qiagen). Total RNA (1 μg) was reverse transcribed to synthesize template cDNA by random hexamers using the SuperScript III First-Strand System (Life Technologies), and 20 μl of synthesized cDNA was diluted fivefold with DEPC-treated water. For RT–PCR, EzWay Taq PCR MasterMix (Komabiotech, Korea) and 5 μl of synthesized cDNA template were used. Thermal cycling was carried out under the following conditions: incubation at 95 °C for 1 min followed by 30 cycles of 30 s at 95 °C, 30 s at 56 °C, and 30 s at 72 °C. The primer pairs used in this experiment were designed to generate the amplification product, including the breakpoints of the fusion genes. PCR products were analyzed by agarose gel electrophoresis. The primers are summarized in **Supplementary Table 6**.

Data availability. All sequenced data have been deposited in the European Genome-phenome Archive (EGA) with accession code EGAS00001001880 (RNA–seq and WES data).

33. Lee, J. *et al.* Tumor stem cells derived from glioblastomas cultured in bFGF and EGF more closely mirror the phenotype and genotype of primary tumors than do serum-cultured cell lines. *Cancer Cell* **9**, 391–403 (2006).
34. Clark, K. *et al.* The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *J. Digit. Imaging* **26**, 1045–1057 (2013).
35. Wangaryattawanich, P. *et al.* Multicenter imaging outcomes study of The Cancer Genome Atlas glioblastoma patient cohort: imaging predictors of overall and progression-free survival. *Neuro-oncol.* **17**, 1525–1537 (2015).
36. Rios Velazquez, E. *et al.* Fully automatic GBM segmentation in the TCGA-GBM dataset: Prognosis and correlation with VASARI features. *Sci. Rep.* **5**, 16822 (2015).
37. Giannopoulos, S. & Kyritsis, A.P. Diagnosis and management of multifocal gliomas. *Oncology* **79**, 306–312 (2010).
38. Kyritsis, A.P., Levin, V.A., Yung, W.K. & Leeds, N.E. Imaging patterns of multifocal gliomas. *Eur. J. Radiol.* **16**, 163–170 (1993).
39. Zinn, P.O. *et al.* Radiogenomic mapping of edema/cellular invasion MRI-phenotypes in glioblastoma multiforme. *PLoS One* **6**, e25451 (2011).
40. Li, Y.M., Suki, D., Hess, K. & Sawaya, R. The influence of maximum safe resection of glioblastoma on survival in 1229 patients: can we do better than gross-total resection? *J. Neurosurg.* **124**, 977–988 (2016).
41. Carter, S.L. *et al.* Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* **30**, 413–421 (2012).
42. Suzuki, H. *et al.* Mutational landscape and clonal architecture in grade II and III gliomas. *Nat. Genet.* **47**, 458–468 (2015).
43. Li, B. & Dewey, C.N. RSEM: accurate transcript quantification from RNA–Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
44. Iyer, M.K., Chinnaiyan, A.M. & Maher, C.A. ChimeraScan: a tool for identifying chimeric transcription in sequencing data. *Bioinformatics* **27**, 2903–2904 (2011).
45. Cahoy, J.D. *et al.* A transcriptome database for astrocytes, neurons, and oligodendrocytes: a new resource for understanding brain development and function. *J. Neurosci.* **28**, 264–278 (2008).
46. Butovsky, O. *et al.* Identification of a unique TGF-β-dependent molecular and functional signature in microglia. *Nat. Neurosci.* **17**, 131–143 (2014).
47. Bhasin, M. *et al.* Bioinformatic identification and characterization of human endothelial cell–restricted genes. *BMC Genomics* **11**, 342 (2010).
48. Chtanova, T. *et al.* Identification of T cell–restricted genes, and signatures for different T cell responses, using a comprehensive collection of microarray datasets. *J. Immunol.* **175**, 7837–7847 (2005).
49. Abbas, A.R. *et al.* Immune response *in silico* (IRIS): immune-specific genes identified from a compendium of microarray expression data. *Genes Immun.* **6**, 319–331 (2005).
50. Huang, S. & Pang, L. Comparing statistical methods for quantifying drug sensitivity based on *in vitro* dose-response assays. *Assay Drug Dev. Technol.* **10**, 88–96 (2012).
51. Zairis, S. *et al.* Moduli spaces of phylogenetic trees describing tumor evolutionary patterns. *Lect. Notes Comput. Sci.* **8609**, 528–539 (2014).

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.

HHS Public Access
Author manuscript
Fertil Steril. Author manuscript; available in PMC 2019 June 01.

Published in final edited form as:
Fertil Steril. 2018 June ; 109(6): 952–963. doi:10.1016/j.fertnstert.2018.05.006.

Personalized Medicine: Motivation, Challenges and Progress

Laura H. Goetz, M.D. and **Nicholas J. Schork, Ph.D**
Scripps Clinic Medical Group, La Jolla, CA (LHG); The Translational Genomics Research Institute (TGen; NJS), Phoenix, AZ; The City of Hope/TGen IMPACT Center (NJS), Duarte, CA; J. Craig Venter Institute, La Jolla, CA (JCVI; NJS); The University of California, San Diego, La Jolla, CA (UCSD; NJS)

Abstract

There is a great deal of hype surrounding the concept of 'personalized' medicine. Personalized medicine is rooted in the belief that since individuals possess nuanced and unique characteristics at the molecular, physiological, environmental exposure and behavioral levels, they may need to have interventions provided to them for diseases they possess that are tailored to these nuanced and unique characteristics. This belief has been verified to some degree through the application of emerging technologies such as DNA sequencing, proteomics, imaging protocols, and wireless health monitoring devices, which have revealed great inter-individual variation in disease processes. In this review, we consider the motivation for personalized medicine, its historical precedents, the emerging technologies that are enabling it, some recent experiences including successes and setbacks, ways of vetting and deploying personalized medicines, and future directions, including potential ways of treating individuals with fertility and sterility issues. We also consider current limitations of personalized medicine. We ultimately argue that since aspects of personalized medicine are rooted in biological realities, personalized medicine practices in certain contexts are likely to be an inevitability, especially as relevant assays and deployment strategies become more efficient and cost-effective.

Keywords

Precision medicine; biomarkers; patient monitoring; genomics

Introduction

The application of emerging, high-throughput, data-intensive biomedical assays, such as DNA sequencing, proteomics, imaging protocols, and wireless monitoring devices, has revealed a great deal of inter-individual variation with respect to the effects of, and mechanisms and factors that contribute to, disease processes. This has raised questions about

---

Address Correspondence to: Nicholas J. Schork, Ph.D., The Translational Genomics Research Institute, 445 North Fifth Street, Phoenix, AZ; 85004, 602-343-8429, nschork@tgen.org.

LHG and NJS have no conflicts to declare with respect to this article.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

the degree to which this inter-individual variation should impact decisions about the optimal way to treat, monitor, or prevent a disease for an individual. In fact, it is now widely believed that the underlying heterogeneity of many disease processes suggests that strategies for treating an individual with a disease, and possibly monitoring or preventing that disease, must be tailored or ‘personalized’ to that individual's unique biochemical, physiological, environmental exposure, and behavioral profile. A number of excellent reviews on personalized medicine have been written, including a growing number of textbooks on the subject meant for medical students and clinicians. It should be noted that although many use the term ‘personalized’ medicine interchangeably with the terms ‘individualized’ and ‘precision’ medicine (as we do here), many have argued that there are some important, though often subtle, distinctions between them.(1, 2))

There are a number of challenges associated with personalized medicines, especially with respect to obtaining their approval for routine use from various regulatory agencies. In addition, there have many issues associated with the broad acceptance of personalized medicines on the part of different health care stakeholders, such as physicians, health care executives, insurance companies, and, ultimately, patients. Almost all of these challenges revolve around a need to prove that personalized medicine strategies simply outperform traditional medicine strategies, especially since many tailored or personalized therapies, such as autologous CAR-T cell transplant therapies for certain types of cancer(3) and mutation-specific medicines such as ivacaftor to treat cystic fibrosis (4, 5), can be very expensive(6). In this review we consider the history and motivation of personalized medicine and provide some context on what personalized medicines strategies have emerged in the last few decades, what limitations are slowing their advance, and what is on the horizon. We also consider strategies for proving that personalized medicine protocols and strategies can outperform traditional medicine protocols and strategies. Importantly, we distinguish examples and challenges associated with personalized disease prevention, personalized health monitoring, and personalized treatment of overt disease.

Archibald Garrod and The Precursors Of Personalized Medicine

There is much in the history of western medicine that anticipates the emergence of personalized medicine. For reasons of brevity, we will not focus on all of these events, but rather only a few that we feel encompass the most basic themes behind personalized medicine. More than a century ago Archibald Garrod, an English physician, began studying in earnest diseases that would later become known as ‘inborn errors of metabolism.’ Garrod studied a number of rare diseases with overt, visible phenotypic manifestations including alkaptonuria, albinism, cystinuria and pentosuria. Of these, his focused work on alkaptonuria led to some notoriety when he observed that some members of families exhibiting alkaptonuria showed measurably outlying values for certain basic biochemical assays, e.g., from urine, relative to the values of family members who did not possess alkaptonuria. This led him to conclude that alkaptonuria was due to a specific ‘altered course of metabolism’ among affected individuals, which was subsequently proven correct.(7) Further, in considering other rare diseases like alkaptonuria, Garrod argued that “…the thought naturally presents itself that these [conditions] are merely extreme examples of variation of chemical behavior which are probably everywhere present in minor degrees and that just as

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

no two individuals of a species are absolutely identical in bodily structure neither are their chemical processes carried out on exactly the same lines”. This more than hints at his belief that, at least with respect to metabolism, humans vary widely and that these differences in metabolism could help explain overt phenotypic differences between individuals, such as their varying susceptibilities to diseases and the ways in which they manifest diseases.(8, 9)

Garrod was working in the backdrop of a great deal of debate about the emerging field of genetics. Although the specific entities we now routinely refer to as genes (i.e., stretches of DNA sequence that code for a protein and related regulatory elements), were unknown to Garrod and his contemporaries, he and others often referred to ‘factors’ influencing disease possessed by certain individuals that were consistent with the modern notion of genes. Claims about the very presence of such factors were born out of discussions rooted in the findings of Mendel (later, it would be shown that many of the metabolic outliers Garrod observed in people with diseases like alkaptonuria were due to defects in genes possessed by people with those diseases). Mendel observed consistent connections between the emergence of very specific phenotypes only when certain breeding protocols were followed in peas that anticipated the modern field of genetics.(10) Essentially, as discussed in an excellent book by William Provine,(11) many in the research community at the time debated how genes or factors of the type Garrod and others were considering could explain the broad variation in phenotypic expression observed in nature. One group of academics and researchers, referred to as the ‘Mendelians’ in the historical literature, which included William Bateson and Hugo de Vries, focused on the discrete nature of the factors likely to be responsible for many observable inheritance patterns (such as those of focus in Mendel's studies and observations like Garrod's in the context of rare disease). In opposition to the ‘Mendelians’ were the ‘Biometricians,’ represented most notably by Karl Pearson, whose focus on continuous or graded phenotypes, like height, gave them concerns about how to reconcile such continuous variation with the overtly discrete (‘either/or’) factors and inheritance patterns considered by the Mendelians and researchers like Garrod.

The Mendelian vs. Biometrician debate was resolved to a great extent by the statistician Ronald Fisher in a series of seminal papers. Fisher argued that one could reconcile continuous phenotypic variation with discrete, heritable factors that contribute to this variation by suggesting that many factors (i.e., genes) might contribute in a small way to a particular phenotype. The collective effect, or sum total, of these factors could then create variation in phenotypes that give the appearance of continuity in the population at large (e.g., an individual who inherited only 1 of 25 genetic variants known to increase height would be shorter on average than someone who inherited 10 or 12, and much shorter, relatively speaking, than an individual who inherited 22 or 25).(12) The belief that there might be many genes that contribute to phenotypic expression broadly, some with more pronounced effects and some with less pronounced effects, that interact and collectively contribute to a phenotype in a myriad of ways, has been validated through the application of modern high-throughput genetic technologies such as genotyping chips and DNA sequencing. As a result, much of the contemporary focus on personalized medicine is rooted in the findings of genetic studies, as it has been shown that individuals do in fact vary widely as each individual possesses subsets of literally many millions of genetic variants that exist in the human population as a whole. In addition, subsets of these genetic variants may have arisen

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

as *de novo* mutations and hence may be unique to an individual. These extreme genetic variation explains, in part, why individuals vary so much with respect to phenotypes, in particular their susceptibilities to disease and their responses to interventions.(13) It should be emphasized that although personalized medicine has its roots in the results of genetic studies, it is widely accepted that other factors, e.g., environmental exposures, developmental phenomena and epigenetic changes, and behaviors, all need to be taken into account when determining the optimal way to treat an individual patient (see Figure 1).(14-16)

Another, sadly more obscure, publication was also prescient for personalized medicine, although this publication bears more on the need for clinical practices that are consistent with personalized medicine – rather than a scientific justification of personalized medicine. More than 60 years ago Hogben and Sim considered how clinical practice needs to pay attention to nuanced characteristics of patients in order to determine an appropriate intervention for them.(17-19) Although more will be discussed about their paper in the section on ‘Testing Personalized Medicines,’ suffice it to say that the authors believed that in order to determine an optimal course of action for an individual patient in the absence of any *a priori* understanding of how best to treat that patient given his or her characteristics or profile, a number of items would need to be obtained. Thus, greater information about that patient would have to be gathered, a plan to vet the utility of an intervention chosen on the basis of that information would have to be pursued, and a strategy for incorporating the results of the patient-oriented study into future care would have to be crafted. Although simple in theory, the practical issues surrounding gathering more information about a patient and pursuing an the empirical assessment of a personalized intervention can be daunting. For example, questions surrounding how one can know that a chosen intervention works unless meticulous patient follow-up information is kept, how one would know if a patient satisfied with what they are experiencing with the intervention, and how one could assess the difference between other interventions that could have been chosen and the chosen personalized intervention, would all need to be addressed. In fact, practical issues surrounding the implementation of personalized medicine that Hogben and Sim considered are often overlooked in contemporary discussions about personalized medicine, especially since different technologies for profiling patients are constantly being developed and refined, and more and more evidence for inter-individual variation in factors associated with diseases (from technologies such as DNA sequencing, proteomics, sophisticated imaging protocols, etc.) is emerging.

Early Examples of Personalized Medicine

There have been a great many examples of interventions tailored to individual patient profiles, virtually all of them based on genetic profiles. Before providing a few classic examples, it should be emphasized that personalized medicine can be practiced not only for the treatment of disease, but also for the early detection and prevention of disease. We provide some historical examples of personalized disease treatments here and consider early detection and prevention in the next section, as developments in personalized disease detection and prevention are much more recent.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

The human body deals with traditional pharmacotherapies (i.e., drugs) to treat disease in two general ways. Initially, the body must respond to a drug. This response occurs in steps, with the first step involving the absorption of the drug by the body. The drug must then be distributed throughout the body (during this process the drug might be ‘biotransformed’ or metabolized into useful components) and then begin to elicit effects. Finally, any remaining drug or drug components are excreted. These processes are often lumped under the heading of ‘pharmacokinetics’ and collectively referred to as the ‘ADME’ of a drug (Absorption, Distribution, Metabolism and Excretion). Pharmacokinetic activity is often under the control of a unique set of genes (e.g., drug metabolizing enzymes) that could harbor naturally-occurring genetic variants (or ‘polymorphisms’) that influence their function and hence how the body ultimately deals with a particular drug. Once a drug is within the body, how it interacts with its target (typically a gene or protein encoded by a gene) to elicit an effect is known as its ‘pharmacodynamic’ properties. These properties include the ‘affinity’ the drug has for its target(s), the drug's ability to modulate the target(s) (or its ‘efficacy’), and the ‘potency’ of the drug, or how much of the drug is needed to induce a certain change in its target. Pharmacodynamic properties of a drug are also under genetic control.

Many early examples of personalized medicines were associated with genetically-mediated pharmacokinetic aspects of drugs. This was due in part to the biomedical science community's understanding of drug metabolizing enzymes and the role they play in the body's response to drugs. An excellent introduction to pharmacogenetic properties of drugs, as well genetic variants in genes that influence the efficacy and side effects of drugs (especially with respect to genetic variants in drug metabolizing enzymes) is the book by Weber.(20) Warfarin is a widely used blood thinning medication that, if not dosed properly, could cause a potentially life-threatening adverse drug reaction. Warfarin targets a particular gene, VKORC1, and is metabolized in part by the gene CYP2C9. Naturally-occurring genetic variation in both the VKORC1 and CYP2C9 genes leads to variation in the pharmacodynamic and pharmacokinetic properties of Warfarin across individuals, ultimately creating variation in individuals' responses to warfarin. The US Food and Drug Administration has therefore recommended that dosing for warfarin take into consideration an individual's genotype (i.e., the dose must be personalized to an individual based on the specific genetic variants they possess in the VKORC1 and CYP2C9 genes).(21)

Another classic example of a drug that should only be provided to individuals with a certain genetic profile is primaquine (PQ). PQ has been used to manage malaria with some success in parts of the world where malaria is endemic. However, military doctors working in the past observed that some of the soldiers they treated for malaria that were provided the drug became jaundiced and anemic, and ultimately exhibited symptoms of what would later be termed ‘acute haemolytic anaemia (AHA)’. It was later shown that the individuals exhibiting AHA after PQ administration carried variants in the G6PD gene.(22) Current clinical practice with PQ therefore calls for the genotyping of individual patients to see if they carry relevant variants in the G6PD gene that might discourage PQ use for them.

A final, often-cited example of a personalized medicine is the drug imatinib.(23) Imatinib is used to treat chronic myelogenous leukemia (CML). Imatinib inhibits an enzyme, tyrosine kinase, that is increased by the formation of a fusion of two genomic regions, one

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

encompassing the Abelson proto-oncogene (abl) and the other the breakpoint cluster region (bcr). This fusion event arises in many tumors contributing to the development of CML and is referred to as the ‘bcr-abl fusion’ or ‘Philadelphia chromosome.’ However, not all individuals with CML have tumors harboring the bcr-abl fusion mutation. Therefore, imatinib is typically given only to individual CML patients with this fusion event.

Contemporary Examples Of Personalized Medicine

Drugs like warfarin, PQ and imatinib that appear to only work – or only work without side effects – when a patient possesses a certain genetic profile, have generated tremendous interest in identifying factors, such as genetic variants, that influence an individual patient's response to any number of drugs and interventions. This interest in crafting personalized medicines to treat diseases has expanded into personalized disease surveillance (i.e., early detection protocols) and personalized disease prevention strategies as well. We briefly describe a few very recent examples of this activity.

Mutation-Specific Therapies

Instead of developing a drug and then identifying factors that mitigate its efficacy or side effects through observational studies on individuals provided the drug, as with warfarin, PQ and imatinib, there are now attempts to identify, e.g., genetic profiles possessed by patients and then craft therapies that uniquely target those profiles. For example, the drug ivacaftor mentioned earlier was designed to treat individuals with cystic fibrosis (CF) that have very specific pathogenic mutations in the gene CFTR.(4) The CFTR gene has many functions, but one set of functions is dictated by a ‘gate-like’ structure in the CFTR gene's encoded protein that can open and close to control the movement of salts in and out of cells. If the CFTR gene is dysfunctional, then the gate is closed, causing a build-up of mucus and other material in the lungs. Different mutations in the CFTR gene cause different types of dysfunction. For example, some mutations simply cause the CFTR gene to not produce anything, whether the gate is open or not. Other mutations cause the gate mechanism to dysfunction. Ivacaftor is designed to open the gate for longer periods of time in the presence of certain mutations that tend to cause the gate to be closed. Therefore, ivacaftor is only useful for the small subset of CF patients whose CFTR mutations lead to this specific gating problem. Connections between genetic variants and drug efficacy and side effects are growing in number, and in fact the US FDA provides a list of approved drug-based interventions that require a test to determine their appropriateness for an individual: https://www.fda.gov/Drugs/ScienceResearch/ucm572698.htm. Other publications consider the practical implications of approved personalized medicine interventions, such as the report produced by the Personalized Medicine Coalition (PMC).(24)

A second example involves the emerging set of cancer treatments known as immunotherapies.(25) Although there are many types of immunotherapies, all of them seek to prime or trigger an individual's own immune system to attack a cancer. One type of immunotherapy exploits potentially unique sets of genetic alterations that arise in a cancer patient's tumor cells, known as ‘neo-antigens,’ which are often capable of raising an immune response if recognized properly by the host's immune cells. Essentially, this type of

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

immunotherapy works by harvesting cells from a patient that mediate that patient's immune reactions, such as T cells, then modifying those cells to specifically recognize and target the neo-antigens found to be present in the patient's tumor. These modified cells are then put back in the patient's body so these cells can attack the tumor cells giving off the neo-antigen signals. Cytotoxic T cell therapies like this, as well as immunotherapies in general, have had notable successes, but can be very patient-specific for two reasons. First, the neo-antigen profile of a patient might be very unique, such that cytotoxic T cells made to recognize and attack a specific set of neo-antigens will not work in someone whose tumor does not have those neo-antigens. Second, if ‘autologous’ constructs are used, then the patient's own T cells are modified, and hence not likely to work as well in another patient, although attempts to make ‘allogeneic’ constructs in which one individual's T cells are modified and introduced into another patient's body are being pursued aggressively.(25)

Personalizing Early Detection Strategies

If an individual is susceptible to a disease, or susceptible to recurrence of a disease, then that individual should be monitored. It is now believed that such monitoring should be pursued with use of ‘personal thresholds,’ as opposed to ‘population thresholds,’ to make claims about evidence or signs of disease or a pathogenic process.(26) Population thresholds are derived from epidemiologic data and population surveys and include, for example, cholesterol levels > 200 being an indicator for risk of heart disease, or systolic blood pressure > 140 being an indicator of hypertension, risk of stroke or heart disease. Personal thresholds are established from legacy values of a measure collected on an individual over time that used to gauge how deviant future values of that measure might be for that individual. Significant deviations from historical or average legacy values are taken as a sign of a health status change, irrespective of whether or not those values are beyond a population threshold.(27) As an example, Drescher et al.(26) explored the utility of personal thresholds applied to longitudinal CA125 levels collected on a number of women, a subset of whom developed ovarian cancer. The authors found that in all but one instance, the application of personal thresholds would have captured the presence of ovarian cancer at the same time as, or importantly earlier than, the application of population thresholds. Further, the authors showed that the use of personal thresholds could have captured the ovarian cancer almost a year earlier, on average, then the use of population thresholds. As the costs and convenience associated with monitoring assays and technologies improves (i.e., they become cheap and non-intrusive, if not transparent, to an individual user, say through an easily implantable wireless device), then the use of personal thresholds will likely become the rule rather than the exception in health monitoring protocols.

Personalizing Disease Prevention

The use of genetic information to develop personalized disease prevention strategies is now well established in the scientific community, but not yet widely adopted in clinical practice. There are many excellent examples of how the use of genetic information can lead to both a decreased risk of disease development as well as decreased complications from standard treatment and screening strategies. A prime example relates to colorectal cancer, which remains the third leading cause of cancer deaths despite being a highly preventable illness. In 2012 Liao et al. reported an improvement in overall survival and a decreased risk for

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

cancer-specific deaths in patients taking postoperative aspirin if they exhibited a somatic mutation in the PIK3CA gene in their colorectal cancers compared with patients whose cancers had the wild-type PIK3CA gene.(28) In 2015, Nan et al. reported varying effects of aspirin use on risk for development of colorectal cancer depending on an individual's genotype, with individuals possessing different genotypes having either lower, higher or no change in their risk of colorectal cancer development with aspirin use.(29) Given that aspirin use can have serious side effects associated with intestinal and intracranial bleeding, it would be ideal to limit the use of this medication for those individuals predicted to have a side effect, based on genotype.

As another example, in 2018, Jeon et al. reported the use of expanded risk prediction models for determining when to begin colorectal cancer screening. Currently the guidelines use only age and family history as variables. Jeon et al. showed that by using information about an individual's environmental exposure and genetic profile, specifically the presence of colorectal cancer associated genetic variants, recommendations for when to start screening could change by 12 years for men and 14 years for women.(30) The accuracy of relevant predictions about an individual's risk for colorectal cancer has been studied and suggests that the area under the curve (AUC) value for a model including environmental and genetic factors, where an AUC of 1.0 would suggest a model with perfect predictive accuracy, was 0.63 for men and 0.62 for women. This is compared to an AUC value of 0.53 (men) 0.54 (women) when only family history information was considered. Although there is still room for improvement given the AUCs were only ~0.62 for the model with patient environmental exposure and genetic information, the considerable improvement over models that did not include genetic or environmental information justifies their use.

Testing Personalized Medicines

Although we have argued that personalized medicine is rooted in a great number of legacy insights and historical precedents, mostly related to genetics and rare diseases, its recognition as a paradigm that should be embraced broadly by the biomedical research and clinical communities is relatively recent. This suggests that not enough time has elapsed since the time of this recognition for researchers to show that personalized medicine actually works in a wide enough variety of settings to motivate its broad adoption. In this light, questions of how the community can vet or test the utility of personalized medicine arise. We describe three emerging strategies for vetting personalized medicines below, including N-of-1 clinical trials, intervention-matching trials, and adaptive clinical trials, and argue that although these strategies borrow elements from traditional randomized clinical trials (RCTs), they deviate significantly from historical population-based RCTs that were prominent in the past.

N-of-1 Clinical Trials

If there is no reason to believe that any one of a set of different interventions matches an individual's profile (e.g., genomic, behavioral, etc.) better than others, then there is ‘equipoise’ among those interventions. In this case it becomes an empirical question as to which intervention might be optimal for the individual in question. Trials focusing on an

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

individual's response to different interventions to determine an optimal intervention are referred to as ‘N-of-1’ or single subject trials. N-of-1 trials often exploit a simple cross-over design or even a repeated crossover designs, such as ‘ABABAB’ designs, where ‘A’ and ‘B’ refer to different interventions, and the sequence ‘ABABAB’ refers to the order in which the interventions are provided to a patient. Alternating interventions, and collecting data on the individual's response to those interventions, allows comparisons of those interventions (for example between a test intervention and a comparator, or placebo, intervention. Randomization, blinding, washout periods, multiple endpoints, and many other design elements can be used in N-of-1 trials.(27, 31, 32)

N-of-1 trials involving the provision of different interventions in sequence to an individual and evaluating outcomes for each, need to accommodate serial correlation between the observations, as well as possible carry-over effects from one intervention to another, but these issues can largely be overcome with appropriate analytical methods and study design.(32) Cross-over based N-of-1 trials are impractical, if not unethical, in settings where an individual is suffering from an acute or life-threatening condition, since switching from one intervention to another may exacerbate that individual's condition. However, sequential N-of-1 designs, in which measures are continuously monitored in real time to determine if an intervention is causing harm or working, have been proposed for these situations.(27) Given that the focus of an N-of-1 trial is on the identification of an optimal intervention for an individual, rather than on the average response to an intervention in the population at large (which is often the focus of traditional RCTs), they may be most appropriate to conduct in actual clinical practice when a physician is faced with equipoise, as considered by Hogen and Sim.(33, 34)

Intervention Matching Trials

If evidence is found that certain features in individual patients' profiles can be used to identify interventions that might work for each of them, then a question arises as to how to test that the hypothesis that providing interventions to those individuals based on these ‘matches’ leads to better outcomes than providing those individuals interventions based on some other scheme or strategy. One could test each individual match, but this may require pursuing many small clinical trials, which may be logistically complicated and hard to find financial support and infrastructure to implement. As an alternative, one could test an entire matching strategy against an alternative way of providing interventions (e.g., giving everyone the same intervention). This is more or less the motivation behind ‘basket’ and ‘umbrella’ trials currently in use, primarily in oncology settings.(35, 36) In oncology contexts, basket and umbrella trials enroll multiple individual patients into a trial knowing that they each might have unique features in their profile that could indicate that different interventions are appropriate. Basket trials enroll individuals without regard to the specific tissue affected by cancer (e.g., lung, breast and colorectal cancer patients can be enrolled) whereas umbrella trials only consider a single tissue (only lung cancer patients are enrolled). Each patient's tumor is profiled, usually via DNA sequencing. The tumor genome is analyzed to see if there are actionable ‘driver’ perturbations in the tumor, such as mutations affecting particular genes, that are likely contributing to the growth of the tumor. If the mechanisms of action of a group of interventions (i.e., cancer drugs) are understood well

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

enough, it may be possible to match those drugs to the perturbations in the tumor (e.g., if the EGFR gene is mutated and overexpressed in the tumor, then using a drug like cetuximab, which inhibits the EGFR gene, would be logical). Thus, each patient is steered towards a particular intervention basket (e.g., the EGFR inhibitor basket). The trial then seeks to test the hypothesis that the use of the different intervention baskets based on the matching scheme results in better outcomes than interventions provided to individual patients based on some other scheme that does not involve tumor profiling and matching.

If the trial is a failure (i.e., the matching scheme doesn't lead to better outcomes than something else), then an argument could be made that the matching scheme was flawed and not necessarily that the interventions considered in the trial are flawed. It would also be wrong to assume that the concept of personalized medicine is flawed as a result of a failure of a basket or bucket trial if in fact the matching scheme was found to be flawed. Some basket trials only have a single basket and no comparison group, but rely on determining which patient profiles appear to be associated with better outcomes for the intervention being tested.(37) Intervention matching schemes are likely to become the rule rather than the exception in medicine, especially since the introduction of computational environments like IBM's ‘Watson’ system. Essentially, Watson is system that includes a very large database extracted in part from the vast medical literature, providing links between information about a patient (e.g., genetic profiles, age, sex, etc.) to outcomes (such as drug response). These links have been enhanced by leveraging statistical methods to further assess relationships between patient profiles and outcomes. For example, Watson has been ‘trained’ to identify and establish links about perturbations often observed in a tumor and how those perturbations might be combatted by available drugs and interventions generally. Thus, if Watson was provided a patient profile, it could look up the best possible intervention given the current state of the science reflected in the literature and Watson's methods for establishing links between profiles and outcomes. The use of IBM's Watson system in actual clinical settings has led to discussions about how best to test and deploy such as a system as a way of supporting, as opposed to replacing, physicians' decisions about an intervention choice for individual patients.(38)

Adaptive Clinical Trials

Adaptive and sequential clinical trials have been used for decades but their consideration and use in personalized medicine contexts is much more recent.(35) Essentially, adaptive trials have as one of their focal points a desire to minimize the amount of time a patient is on what is likely to be an inferior therapy. In the context of personalized medicine, if there is equipoise with respect to available interventions or between an untested and a conventional intervention for an individual patient, then the evaluation of the effects of each intervention on an individual to determine the best one for that individual (as in a very elaborate N-of-1 study) might be impractical or cause more harm than good. This is the case because some, if not all, of the interventions might not actually benefit that individual. In this light, it makes sense to implement studies in which biomarkers reflecting response or adverse effects are collected on an individual trial participant and monitoring of those biomarkers is pursued to determine if there are signs an intervention is not working. If there are, e.g., signs that an intervention is not working, the individual could cross-over to a new intervention. Although

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

adaptive designs can be difficult to implement given their real-time evaluation and updating components, and can also produce data that might be more complicated to analyze than data from fixed, non-adaptive trials, they are often seen as more ethical. In addition, adding adaptive components to N-of-1 and aggregated N-of-1 trials as well as intervention-matching trials is possible. Although there are a growing number of papers describing adaptive trials, the work of Murphy and colleagues has received a great deal of attention because of its focus on minimizing the amount of time a patient is on an inferior treatment.(39-41)

Emerging and Next-Generation Personalized Medicine Strategies

There are a number of very recent research and clinical activities that are charting new territory for personalized medicine. We focus on four of these activities in the following, providing a brief overview of each. These activities include the use of patient-derived cell and organoid ‘avatars’ for determining the best therapies for that patient, the use of intense individualized diagnostic and monitoring protocols to detect signs of disease, the development of personalized digital therapeutics, and the use of personalized medicine approaches in treating patients with fertility issues.

Patient-Derived Cellular Avatars

It is now possible to harvest cells from individuals and use pluripotency induction (i.e., induced pluripotent stem cell or ‘iPSC’) methods on those cells to generate additional cell types of relevance to a patient's condition without having to directly biopsy the affected tissue. This allows researchers to essentially develop a ‘disease in a dish’ cellular model of a patient's condition.(42-44) These *in vitro* cellular ‘avatars’ can be studied to identify key molecular pathologies that might give an indication as to how best to treat an individual patient of interest. The use of iPSC technologies in this manner can be extended with a few additional, very recently developed, technologies to create even better models of an individual's condition. For example, if the patient has a known mutation causing his or her condition, it is possible to use assays based on, e.g., Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) and related constructs to create isogenic cells in which some cells have the mutation in question and some do not. Comparison of these cells allows direct insight into the effects of the mutation while controlling for all relevant genetic background effects associated with the patient's genome.(45, 46) In addition, it is possible to create partial organs or ‘organoids’ from cells obtained from an individual.(47) Organoids can provide greater insight into molecular pathologies associated with an individual patient's condition since they can model cell:cell interactions and more global tissue function.(48)

To achieve truly personalized medical care, the use of patient avatars derived from their own cells could be integrated with other pieces of information about a patient, as well as protocols for acting on that information. Schork and Nazor describe the motivation and integration of different aspects of patient diagnosis, intervention choice, and monitoring, using, among other things, patient avatars.(49) One important aspect of the use of cell-based patient avatars in personalized medicine is that they can accommodate personalized drug screening: literally testing thousands of drugs and compounds against a patient's cells (or

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

organoids, possibly modified with CRISPR technologies) to identify drugs or compounds that uniquely correct the patient's molecular defects. If the drug or compound has actually been approved for use, possibly for another condition, then it could be tested for efficacy with the patient in question under an approved drug ‘repurposing’ protocol. The use of patient-derived cells in personalized drug screening initiatives has shown some success in cancer settings, as tumor biopsies can yield appropriate material for drug screening.(48, 50) The biggest concern with this approach revolves around the question of whether or not the *in vitro* models capture relevant *in vivo* pathobiology and drug response information that may impact a patient's response to a chosen drug. A more direct strategy for *in vivo* experimental cancer intervention choice could involve implanting a device into a patient's tumor *in vivo* and then delivering different drugs through that device to see which ones have an effect.(51, 52)

Intensive Personalized Health Monitoring

The availability of inexpensive genotyping and sequencing technologies is allowing individuals and their health care providers to assess their genetically-mediated risk for disease and/or make a genetic diagnosis if they are already diseased. In addition, given the availability of health monitoring devices, online-ordered blood-based clinical assays, inexpensive imaging devices, etc. it is possible to continuously, or near continuously, monitor aspects of an individual's health (see Figure 1 and see the articles associated with the ‘quantified self’ movement: http://quantifiedself.com)(53, 54). With this in mind, combining genetic risk or diagnostic assessment with intense health monitoring makes sense. A number of individuals with unique diseases and conditions have benefitted from a genetic diagnosis, as it uncovered potential genetically-mediated pathogenic mechanisms or revealed potential targets for pharmacotherapies for them.(49) In addition, a number of individuals have monitored their health intensely for the express purpose of identifying signs of a health status changes, some of which might be attributable to genetic susceptibilities.(55) Table 1 lists examples of published studies exploring the utility of genetic assays in generating a diagnosis for individuals with idiopathic conditions (or what have been referred to as ‘diagnostic odysseys’) as well as published studies exploring the utility of near continuous monitoring to identify evidence for a health status change in an individual. Such diagnoses and monitoring are highly personalized by definition.(15, 16, 56)

Monitoring individuals for health status changes is not trivial, however, if the measures being collected have not been evaluated in a population. This is because there will be no established norms that can be contrasted to the measures collected on an individual to know if those measures are abnormal. However, the community is quickly recognizing the utility of establishing ‘personal thresholds’ for measures as opposed to ‘population thresholds,’ as discussed in the ‘Personalizing Early Detection Strategies’ section above (26, 27) As noted, population thresholds are established from epidemiologic and population survey data and include often-used thresholds for determining disease status such as a cholesterol level greater than 200 for heart disease or a systolic blood pressure greater than 140 mmHg for hypertension. Personal thresholds are established from longitudinal or legacy values of a measure collected on an individual and may be unique to the individual in question and their use in some settings suggests that they work better than population thresholds.(26)

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Digital Therapeutics and Personalized App Content

The ubiquity of smart phones has attracted the interest of many researchers in the health professions as a vehicle for not only collecting health data through various ‘apps’ but also to provide advice, feedback, coaching, imagery, music, text-messages, or connections with other resources, that could benefit an individual with a particular condition or disease. This has led to the emergence of the concept of a ‘digital therapeutic:’ a smart phone app designed to treat and bring relief to an individual affected by a medical or psychological condition.(57) The content provided by a digital therapeutic app to an individual could vary depending on what is learned about that individual and his or her response to content provided in the app. In this way, the app can be personalized.(58) Many digital therapeutics have undergone evaluation for their ability to engage users and benefit them.(59) The US Food and Drug Administration (FDA) has created guidelines for registering digital therapeutics as bona-fide, insurance-reimbursable, approved health technologies, and has begun evaluating and approving many of them. The first approved digital therapeutic – an app for substance abuse – was approved by the FDA in 2017.(60) How easily digital therapeutics will be assimilated into the care stream is an open question.(61)

Personalized Interventions Involving Fertility and Sterility

Personalized medicine strategies and approaches can be applied to treatments for fertility, as many researchers have proposed. For example, it has been suggested that one could leverage ‘real world’ data of the type collected routinely on patients visiting reproductive medicine and fertility clinics (from, e.g., Electronic Medical Record (EMR) systems established at many hospitals and clinics), and use these data to in analyses exploring patterns, patient subgroups and individual patient profiles that could shed light on variation in fertility rates, responses to interventions to enhance fertility, etc. The results of these analyses could then guide future care for patients with fertility issues.(62) In the context of the use of digital medicine, proposals to develop smart phone apps that could provide personalized coaching content to enhance pregnancy have been put forth.(63). Genetic variants known to influence fertility have also been identified and could be used to support diagnoses or personalized intervention plans.(64)(65) Finally, adaptive trial designs have been proposed that could be used to assess the utility of personalized approaches to raising awareness about time to conception and fertility.(66)

In addition to these more traditional approaches to personalizing fertility interventions, there are a number of emerging strategies to enhance fertility in women that go beyond traditional ways of stimulating ovaries.(67) For example, it is now possible to cryopreserve a set oocytes and ovarian tissue samples from a woman and then implant them in her at a later time that may suit her desire to become pregnant.(68) Such a procedure would be highly personalized, since it would work with an individual's own cells and accommodate her preferences for becoming pregnant. However, this procedure would only work if the preserved tissues were viable and not damaged, although relevant cells in those tissues could, in theory, be corrected for genetic defects using gene editing techniques.(69) A more futuristic and controversial personalized fertility intervention, involves the concept that one could use cell reprogramming technologies to generate sperm and egg cells from other cells

Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

obtained from an individual (e.g., skin cells) that could be edited to generate *de novo* gametes for fertilization – a concept known as '*in vitro* gametogenesis.'(70)

Conclusions

Personalized Medicine, or the practice of characterizing an individual patient on a number of levels (e.g., genomic, biochemical, behavioral, etc.) that might shed light on their response to an intervention, and then treating them accordingly, is a necessity given the fact that clinically meaningful inter-individual variation has, and will continue to be, identified. The availability of modern biomedical technologies such as DNA sequencing, proteomics, and wireless monitoring devices, has enabled the identification of this variation, essentially exposing the need for the personalization of medicine at some level. The future challenges associated with this reality will be to not only improve the efficiency in the way in which individuals are characterized, but also in the way in personalized medicines are crafted and vetted to show their utility. This is not to say that interventions that work ubiquitously (i.e., the traditional single agent 'block buster' drugs) should be ignored if identified, but rather that they might be very hard to identify going forward.

There are a few other issues associated with personalized medicine that may hard to overcome in the near term. For example, the need for large data collections in order to identify factors that discriminate groups of individuals that might benefit more from one or another intervention, could create concerns about privacy and the data about those individuals possibly being used for nefarious purposes.(71-73) Fortunately, this issue is not necessarily unique to health care settings, whether current or future, as it has plagued many other industries including the banking, marketing, and social media industries. Strategies exploited in these other industries could be used in health care settings as well. In addition, developing more efficient ways of developing personalized medicines (for example, with respect to cell replacement therapies or mutation-specific drugs that work for a small fraction of patients) is crucial to meet the demands of all patients. Also, paying for personalized medicine practices in the future may be complicated given that they might be initially more expensive.(74) Finally, in order for various stakeholders to embrace personalized medicine, better strategies to educate and train health care professionals about personalized medicine must be developed and implemented.

Acknowledgments

Dr. Schork and his lab are funded in part by US National Institutes of Health Grants UL1TR001442 (CTSA), U24AG051129, U19G023122, as well as a contract from the Allen Institute for Brain Science (note that the content of this manuscript is solely the responsibility of the authors and does not necessarily represent the official views of the NIH).

References

1. CDC. 2016. Available from: https://blogs.cdc.gov/genomics/2016/04/21/shift/
2. NIH. 2018. Available from: https://ghr.nlm.nih.gov/primer/precisionmedicine/precisionvspersonalized
3. Shah GL, Majhail N, Khera N, Giralt S. Value-Based Care in Hematopoietic Cell Transplantation and Cellular Therapy: Challenges and Opportunities. Curr Hematol Malig Rep. 2018; doi: 10.1007/s11899-018-0444-z

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

4. Davis PB, Yasothan U, Kirkpatrick P. Ivacaftor. Nat Rev Drug Discov. 2012; 11(5):349–50. DOI: 10.1038/nrd3723 [PubMed: 22543461]

5. Gulland A. Cystic fibrosis drug is not cost effective, says NICE. BMJ. 2016; 353:i3409.doi: 10.1136/bmj.i3409 [PubMed: 27325384]

6. Check, Hayden E. Promising gene therapies pose million-dollar conundrum. Nature. 2016; 534(7607):305–6. DOI: 10.1038/534305a [PubMed: 27306167]

7. Garrod AE. The Incidence of Alkaptonuria: A Study of Chemical Individuality. Lancet. 1902; 160(4137):1616–20.

8. Garrod, AE. Inborn Errors of Metabolism 2nd ed London: Henry Frowde and. Hodder & Stoughton; 1923.

9. Garrod, AE. The Inborn Factors in Disease: An Essay. Oxford: Clarendon Press; 1931.

10. Mendel JG. Versuche uber Pflanzenhybriden. Verhandlungen des naturforschenden Vereines in Brünn. 1865:3–47.

11. Provine, WB. The Origins of Theoretical Population Genetics. Chicago: University of Chicago Press; 1971.

12. Fisher RA. The Correlation between Relatives on the Supposition of Mendelian Inheritance. Philosophical Transactions of the Royal Society of Edinburgh. 1918; 52:399–433.

13. Ginsberg, GS, Willard, HF, editorsGenomic and Personalized Medicine. 2nd. London: Academic Press; 2013.

14. Carlsten C, Brauer M, Brinkman F, Brook J, Daley D, McNagny K, et al. Genes, the environment and personalized medicine: We need to harness both environmental and genetic data to maximize personal and population health. EMBO Rep. 2014; 15(7):736–9. DOI: 10.15252/embr.201438480 [PubMed: 24907240]

15. Schork NJ. Genetic parts to a preventive medicine whole. Genome Med. 2013; 5(6):54.doi: 10.1186/gm458 [PubMed: 23806045]

16. Patel CJ, Sivadas A, Tabassum R, Preeprem T, Zhao J, Arafat D, et al. Whole genome sequencing in support of wellness and health maintenance. Genome Med. 2013; 5(6):58.doi: 10.1186/gm462 [PubMed: 23806097]

17. Hogben L, Sim M. The self-controlled and self-recorded clinical trial for low-grade morbidity. Br J Prev Soc Med. 1953; 7(4):163–79. [PubMed: 13106260]

18. Glasziou PP. Commentary: the history and place of n-of-1 trials: a commentary on Hogben and Sim. Int J Epidemiol. 2011; 40(6):1458–60. DOI: 10.1093/ije/dyr031 [PubMed: 22033294]

19. Tabery J. Commentary: Hogben vs the Tyranny of averages. Int J Epidemiol. 2011; 40(6):1454–8. DOI: 10.1093/ije/dyr027 [PubMed: 22158660]

20. Weber, WW. Pharmacogenetics. Oxford, editorOxford: Oxford University Press; 1997.

21. Lee MT, Klein TE. Pharmacogenetics of warfarin: challenges and opportunities. J Hum Genet. 2013; 58(6):334–8. DOI: 10.1038/jhg.2013.40 [PubMed: 23657428]

22. Luzzatto L, Seneca E. G6PD deficiency: a classic example of pharmacogenetics with on-going clinical implications. Br J Haematol. 2014; 164(4):469–80. DOI: 10.1111/bjh.12665 [PubMed: 24372186]

23. O'Dwyer ME, Druker BJ. Status of bcr-abl tyrosine kinase inhibitors in chronic myelogenous leukemia. Curr Opin Oncol. 2000; 12(6):594–7. [PubMed: 11085460]

24. PMC. Personalized Medicine at the FDA: 2017 Progress Report. Personalized Medicine Coalition. 2018

25. Farkona S, Diamandis EP, Blasutig IM. Cancer immunotherapy: the beginning of the end of cancer? BMC Med. 2016; 14:73.doi: 10.1186/s12916-016-0623-5 [PubMed: 27151159]

26. Drescher CW, Shah C, Thorpe J, O'Briant K, Anderson GL, Berg CD, et al. Longitudinal screening algorithm that incorporates change over time in CA125 levels identifies ovarian cancer earlier than a single-threshold rule. J Clin Oncol. 2013; 31(3):387–92. DOI: 10.1200/JCO.2012.43.6691 [PubMed: 23248253]

27. Schork NJ, Goetz LH. Single-Subject Studies in Translational Nutrition Research. Annu Rev Nutr. 2017; 37:395–422. DOI: 10.1146/annurev-nutr-071816-064717 [PubMed: 28715990]

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

28. Liao X, Lochhead P, Nishihara R, Morikawa T, Kuchiba A, Yamauchi M, et al. Aspirin use, tumor PIK3CA mutation, and colorectal-cancer survival. N Engl J Med. 2012; 367(17):1596–606. DOI: 10.1056/NEJMoa1207756 [PubMed: 23094721]
29. Nan H, Hutter CM, Lin Y, Jacobs EJ, Ulrich CM, White E, et al. Association of aspirin and NSAID use with risk of colorectal cancer according to genetic variants. JAMA. 2015; 313(11):1133–42. DOI: 10.1001/jama.2015.1815 [PubMed: 25781442]
30. Jeon J, Du M, Schoen RE, Hoffmeister M, Newcomb PA, Berndt SI, et al. Colorectal Transdisciplinary S, Genetics, Epidemiology of Colorectal Cancer C. Determining Risk of Colorectal Cancer and Starting Age of Screening Based on Lifestyle, Environmental, and Genetic Factors. Gastroenterology. 2018; doi: 10.1053/j.gastro.2018.02.021
31. Schork NJ. Personalized medicine: Time for one-person trials. Nature. 2015; 520(7549):609–11. DOI: 10.1038/520609a [PubMed: 25925459]
32. Lillie EO, Patay B, Diamant J, Issell B, Topol EJ, Schork NJ. The n-of-1 clinical trial: the ultimate strategy for individualizing medicine? Per Med. 2011; 8(2):161–73. DOI: 10.2217/pme.11.7 [PubMed: 21695041]
33. Duan N, Kravitz RL, Schmid CH. Single-patient (n-of-1) trials: a pragmatic clinical decision methodology for patient-centered comparative effectiveness research. J Clin Epidemiol. 2013; 66(8 Suppl):S21–8. DOI: 10.1016/j.jclinepi.2013.04.006 [PubMed: 23849149]
34. Kravitz, RL, Duan, N, Eslick, I, Gabler, NB, Kaplan, HC, Larson, EB. , et al. Design and Implementation of N-of-1 Trials: A User's Guide. Rockville, MD: 2014.
35. Biankin AV, Piantadosi S, Hollingsworth SJ. Patient-centric trials for therapeutic development in precision oncology. Nature. 2015; 526(7573):361–70. DOI: 10.1038/nature15819 [PubMed: 26469047]
36. Simon R, Roychowdhury S. Implementing personalized cancer genomics in clinical trials. Nat Rev Drug Discov. 2013; 12(5):358–69. DOI: 10.1038/nrd3979 [PubMed: 23629504]
37. Hyman DM, Piha-Paul SA, Won H, Rodon J, Saura C, Shapiro GI, et al. HER kinase inhibition in patients with HER2- and HER3-mutant cancers. Nature. 2018; 554(7691):189–94. DOI: 10.1038/nature25475 [PubMed: 29420467]
38. Chen Y, Elenee Argentinis JD, Weber G. IBM Watson: How Cognitive Computing Can Be Applied to Big Data Challenges in Life Sciences Research. Clin Ther. 2016; 38(4):688–701. DOI: 10.1016/j.clinthera.2015.12.001 [PubMed: 27130797]
39. Klasnja P, Hekler EB, Shiffman S, Boruvka A, Almirall D, Tewari A, et al. Microrandomized trials: An experimental design for developing just-in-time adaptive interventions. Health Psychol. 2015; 34S:1220–8. DOI: 10.1037/hea0000305 [PubMed: 26651463]
40. Laber EB, Lizotte DJ, Qian M, Pelham WE, Murphy SA. Dynamic treatment regimes: technical challenges and applications. Electron J Stat. 2014; 8(1):1225–72. [PubMed: 25356091]
41. Chakraborty B, Murphy SA. Dynamic Treatment Regimes. Annu Rev Stat Appl. 2014; 1:447–64. DOI: 10.1146/annurev-statistics-022513-115553 [PubMed: 25401119]
42. Beltrao-Braga PC, Pignatari GC, Russo FB, Fernandes IR, Muotri AR. In-a-dish: induced pluripotent stem cells as a novel model for human diseases. Cytometry A. 2013; 83(1):11–7. DOI: 10.1002/cyto.a.22231 [PubMed: 23281003]
43. Sayed N, Liu C, Wu JC. Translation of Human-Induced Pluripotent Stem Cells: From Clinical Trial in a Dish to Precision Medicine. J Am Coll Cardiol. 2016; 67(18):2161–76. DOI: 10.1016/j.jacc.2016.01.083 [PubMed: 27151349]
44. Wu J, Izpisua Belmonte JC. Stem Cells: A Renaissance in Human Biology Research. Cell. 2016; 165(7):1572–85. DOI: 10.1016/j.cell.2016.05.043 [PubMed: 27315475]
45. Song M, Ramakrishna S. Genome Editing in Stem Cells for Disease Therapeutics. Mol Biotechnol. 2018; 60(4):329–38. DOI: 10.1007/s12033-018-0072-9 [PubMed: 29516417]
46. Uppada V, Gokara M, Rasineni GK. Diagnosis and therapy with CRISPR advanced CRISPR based tools for point of care diagnostics and early therapies. Gene. 2018; doi: 10.1016/j.gene.2018.02.066
47. Ho BX, Pek NMQ, Soh BS. Disease Modeling Using 3D Organoids Derived from Human Induced Pluripotent Stem Cells. Int J Mol Sci. 2018; 19(4)doi: 10.3390/ijms19040936

48. Aboulkheyr Es H, Montazeri L, Aref AR, Vosough M, Baharvand H. Personalized Cancer Medicine: An Organoid Approach. Trends Biotechnol. 2018; 36(4):358–71. DOI: 10.1016/j.tibtech.2017.12.005 [PubMed: 29366522]
49. Schork NJ, Nazor K. Integrated Genomic Medicine: A Paradigm for Rare Diseases and Beyond. Adv Genet. 2017; 97:81–113. DOI: 10.1016/bs.adgen.2017.06.001 [PubMed: 28838357]
50. Crystal AS, Shaw AT, Sequist LV, Friboulet L, Niederst MJ, Lockerman EL, et al. Patient-derived models of acquired resistance can identify effective drug combinations for cancer. Science. 2014; 346(6216):1480–6. DOI: 10.1126/science.1254721 [PubMed: 25394791]
51. Jonas O, Landry HM, Fuller JE, Santini JT Jr, Baselga J, Tepper RI, et al. An implantable microdevice to perform high-throughput in vivo drug sensitivity testing in tumors. Sci Transl Med. 2015; 7(284):284ra57.doi: 10.1126/scitranslmed.3010564
52. Klinghoffer RA, Bahrami SB, Hatton BA, Frazier JP, Moreno-Gonzalez A, Strand AD, et al. A technology platform to assess multiple cancer agents simultaneously within a patient's tumor. Sci Transl Med. 2015; 7(284):284ra58.doi: 10.1126/scitranslmed.aaa7489
53. Appelboom G, Camacho E, Abraham ME, Bruce SS, Dumont EL, Zacharia BE, et al. Smart wearable body sensors for patient self-assessment and monitoring. Arch Public Health. 2014; 72(1):28.doi: 10.1186/2049-3258-72-28 [PubMed: 25232478]
54. Swan M. The Quantified Self: Fundamental Disruption in Big Data Science and Biological Discovery. Big Data. 2013; 1(2):85–99. DOI: 10.1089/big.2012.0002 [PubMed: 27442063]
55. Bai J, Shen L, Sun H, Shen B. Physiological Informatics: Collection and Analyses of Data from Wearable Sensors and Smartphone for Healthcare. Adv Exp Med Biol. 2017; 1028:17–37. DOI: 10.1007/978-981-10-6041-0_2 [PubMed: 29058214]
56. Li X, Dunn J, Salins D, Zhou G, Zhou W, Schussler-Fiorenza Rose SM, et al. Digital Health: Tracking Physiomes and Activity Using Wearable Biosensors Reveals Useful Health-Related Information. PLoS Biol. 2017; 15(1):e2001402.doi: 10.1371/journal.pbio.2001402 [PubMed: 28081144]
57. Sverdlov O, van Dam J, Hannesdottir K, Thornton-Wells T. Digital Therapeutics: An Integral Component of Digital Innovation in Drug Development. Clin Pharmacol Ther. 2018; doi: 10.1002/cpt.1036
58. Kaner EF, Beyer FR, Garnett C, Crane D, Brown J, Muirhead C, et al. Personalised digital interventions for reducing hazardous and harmful alcohol consumption in community-dwelling populations. Cochrane Database Syst Rev. 2017; 9:CD011479.doi: 10.1002/14651858.CD011479.pub2 [PubMed: 28944453]
59. Iacoviello BM, Steinerman JR, Klein DB, Silver TL, Berger AG, Luo SX, et al. Clickotine, A Personalized Smartphone App for Smoking Cessation: Initial Evaluation. JMIR Mhealth Uhealth. 2017; 5(4):e56.doi: 10.2196/mhealth.7226 [PubMed: 28442453]
60. FDA. 2017. Available from: https://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm576087.htm
61. Phillips KA, Douglas MP, Trosman JR, Marshall DA. “What Goes Around Comes Around”: Lessons Learned from Economic Evaluations of Personalized Medicine Applied to Digital Medicine. Value Health. 2017; 20(1):47–53. DOI: 10.1016/j.jval.2016.08.736 [PubMed: 28212968]
62. Jungheim ES, Carson KR. Leveraging real-world data to move toward more personalized fertility treatment. Fertil Steril. 2018; 109(4):608–9. DOI: 10.1016/j.fertnstert.2018.01.036 [PubMed: 29653708]
63. van Dijk MR, Koster MPH, Willemsen SP, Huijgen NA, Laven JSE, Steegers-Theunissen RPM. Healthy preconception nutrition and lifestyle using personalized mobile health coaching is associated with enhanced pregnancy chance. Reprod Biomed Online. 2017; 35(4):453–60. DOI: 10.1016/j.rbmo.2017.06.014 [PubMed: 28688924]
64. Yurttas Beim P, Parfitt DE, Tan L, Sugarman EA, Hu-Seliger T, Clementi C, et al. At the dawn of personalized reproductive medicine: opportunities and challenges with incorporating multigene panel testing into fertility care. J Assist Reprod Genet. 2017; 34(12):1573–6. DOI: 10.1007/s10815-017-1068-2 [PubMed: 29063499]

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

65. DeAngelis AM, Roy-O'Reilly M, Rodriguez A. Genetic alterations affecting cholesterol metabolism and human fertility. Biol Reprod. 2014; 91(5):117.doi: 10.1095/biolreprod.114.119883 [PubMed: 25122065]
66. Laber EB, Zhao YQ, Regh T, Davidian M, Tsiatis A, Stanford JB, et al. Using pilot data to size a two-arm randomized trial to find a nearly optimal personalized treatment strategy. Stat Med. 2016; 35(8):1245–56. DOI: 10.1002/sim.6783 [PubMed: 26506890]
67. Jungheim ES, Meyer MF, Broughton DE. Best practices for controlled ovarian stimulation in in vitro fertilization. Semin Reprod Med. 2015; 33(2):77–82. DOI: 10.1055/s-0035-1546424 [PubMed: 25734345]
68. Tao T, Del Valle A. Human oocyte and ovarian tissue cryopreservation and its application. J Assist Reprod Genet. 2008; 25(7):287–96. DOI: 10.1007/s10815-008-9236-z [PubMed: 18670872]
69. Geel TM, Ruiters MHJ, Cool RH, Halby L, Voshart DC, Andrade Ruiz L, et al. The past and presence of gene targeting: from chemicals and DNA via proteins to RNA. Philos Trans R Soc Lond B Biol Sci. 2018; 373(1748)doi: 10.1098/rstb.2017.0077
70. Nagamatsu G, Hayashi K. Stem cells, in vitro gametogenesis and male fertility. Reproduction. 2017; 154(6):F79–F91. DOI: 10.1530/REP-17-0510 [PubMed: 29133304]
71. Mooney SJ, Pejaver V. Big Data in Public Health: Terminology, Machine Learning, and Privacy. Annu Rev Public Health. 2017; doi: 10.1146/annurev-publhealth-040617-014208
72. Shen H, Ma J. Privacy Challenges of Genomic Big Data. Adv Exp Med Biol. 2017; 1028:139–48. DOI: 10.1007/978-981-10-6041-0_8 [PubMed: 29058220]
73. Vayena E, Blasimme A. Biomedical Big Data: New Models of Control Over Access, Use and Governance. J Bioeth Inq. 2017; 14(4):501–13. DOI: 10.1007/s11673-017-9809-6 [PubMed: 28983835]
74. Hughes DA. Economics of Pharmacogenetic-Guided Treatments: Underwhelming or Overstated? Clin Pharmacol Ther. 2018; doi: 10.1002/cpt.1030
75. Worthey EA, Mayer AN, Syverson GD, Helbling D, Bonacci BB, Decker B, Serpe JM, Dasu T, Tschannen MR, Veith RL, Basehore MJ, Broeckel U, Tomita-Mitchell A, Arca MJ, Casper JT, Margolis DA, Bick DP, Hessner MJ, Routes JM, Verbsky JW, Jacob HJ, Dimmock DP. Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. Genet Med. 2011; 13(3):255–62. DOI: 10.1097/GIM.0b013e3182088158 [PubMed: 21173700]
76. Bainbridge MN, Wiszniewski W, Murdock DR, Friedman J, Gonzaga-Jauregui C, Newsham I, Reid JG, Fink JK, Morgan MB, Gingras MC, Muzny DM, Hoang LD, Yousaf S, Lupski JR, Gibbs RA. Whole-genome sequencing for optimized patient management. Sci Transl Med. 2011; 3(87): 87re3.doi: 10.1126/scitranslmed.3002243
77. O'Rawe JA, Fang H, Rynearson S, Robison R, Kiruluta ES, Higgins G, Eilbeck K, Reese MG, Lyon GJ. Integrating precision medicine in the study and clinical treatment of a severely mentally ill person. PeerJ. 2013; 1:e177.doi: 10.7717/peerj.177 [PubMed: 24109560]
78. Chen YZ, Friedman JR, Chen DH, Chan GC, Bloss CS, Hisama FM, Topol SE, Carson AR, Pham PH, Bonkowski ES, Scott ER, Lee JK, Zhang G, Oliveira G, Xu J, Scott-Van Zeeland AA, Chen Q, Levy S, Topol EJ, Storm D, Swanson PD, Bird TD, Schork NJ, Raskind WH, Torkamani A. Gain-of-function ADCY5 mutations in familial dyskinesia with facial myokymia. Ann Neurol. 2014; 75(4):542–9. DOI: 10.1002/ana.24119 [PubMed: 24700542]
79. Wartman LD. A case of me: clinical cancer sequencing and the future of precision medicine. Cold Spring Harb Mol Case Stud. 2015; 1(1):a000349.doi: 10.1101/mcs.a000349 [PubMed: 27148564]
80. Chen R, Mias GI, Li-Pook-Than J, Jiang L, Lam HY, Chen R, Miriami E, Karczewski KJ, Hariharan M, Dewey FE, Cheng Y, Clark MJ, Im H, Habegger L, Balasubramanian S, O'Huallachain M, Dudley JT, Hillenmeyer S, Haraksingh R, Sharon D, Euskirchen G, Lacroute P, Bettinger K, Boyle AP, Kasowski M, Grubert F, Seki S, Garcia M, Whirl-Carrillo M, Gallardo M, Blasco MA, Greenberg PL, Snyder P, Klein TE, Altman RB, Butte AJ, Ashley EA, Gerstein M, Nadeau KC, Tang H, Snyder M. Personal omics profiling reveals dynamic molecular and medical phenotypes. Cell. 2012; 148(6):1293–307. DOI: 10.1016/j.cell.2012.02.009 [PubMed: 22424236]
81. Smarr L. Quantifying your body: a how-to guide from a systems biology perspective. Biotechnol J. 2012; 7(8):980–91. DOI: 10.1002/biot.201100495 [PubMed: 22887886]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

82. David LA, Materna AC, Friedman J, Campos-Baptista MI, Blackburn MC, Perrotta A, Erdman SE, Alm EJ. Host lifestyle affects human microbiota on daily timescales. Genome Biol. 2014; 15(7):R89.doi: 10.1186/gb-2014-15-7-r89 [PubMed: 25146375]

83. Forsdyke DR. Summertime dosage-dependent hypersensitivity to an angiotensin II receptor blocker. BMC Res Notes. 2015; 8:227.doi: 10.1186/s13104-015-1215-8 [PubMed: 26055103]

84. Trammell SA, Schmidt MS, Weidemann BJ, Redpath P, Jaksch F, Dellinger RW, Li Z, Abel ED, Migaud ME, Brenner C. Nicotinamide riboside is uniquely and orally bioavailable in mice and humans. Nat Commun. 2016; 7:12948.doi: 10.1038/ncomms12948 [PubMed: 27721479]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

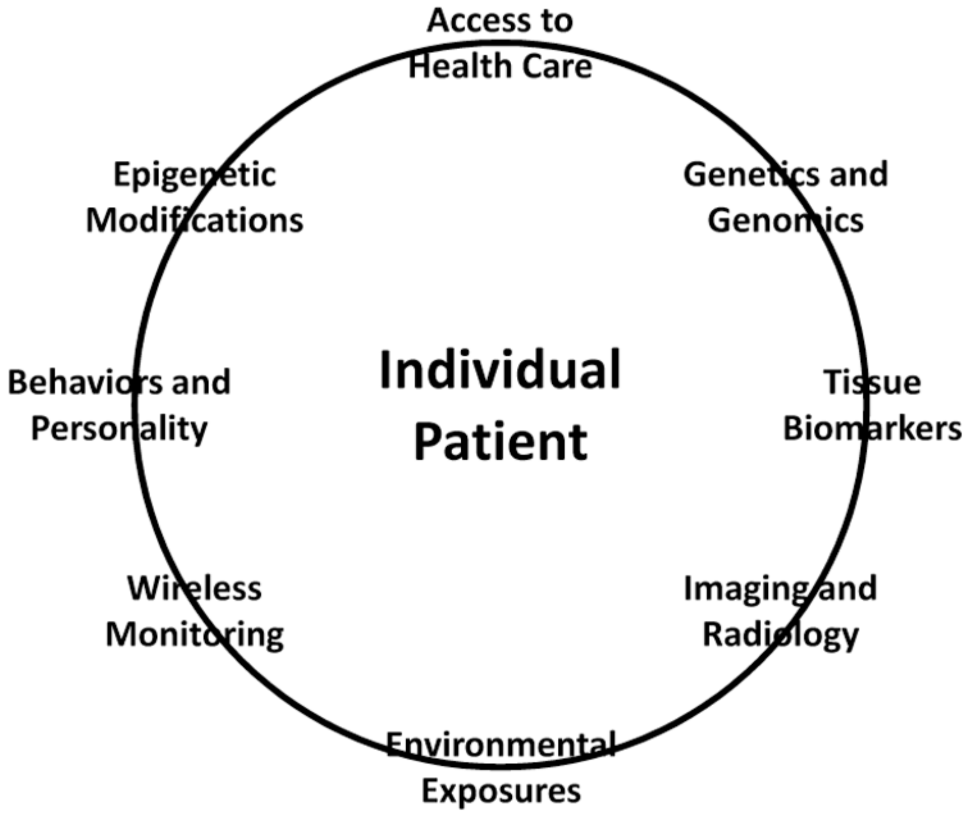

Access to Health Care
Genetics and Genomics
Tissue Biomarkers
Imaging and Radiology
Environmental Exposures
Wireless Monitoring
Behaviors and Personality
Epigenetic Modifications
Individual Patient

Figure 1.

Graphical depiction of elements in need of integration and assessment in pursuing truly personalized medicine. Access to health care is important since some individuals may not be able to access expertise and technologies due to geographic or economic barriers and therefore interventions might need to be crafted for those individuals with this in mind. Inherited genetic information is really only predictive or diagnostic in nature however somatic changes to DNA can provide valuable insight into pathogenic processes. Tissue biomarkers (e.g., routine blood-based clinical chemistry panels) are useful for detecting changes in health status, as are imaging and radiology exams as well as data collected routinely via wireless monitors. Environmental exposures and behaviors can really impact the success of an intervention and exhibit great inter-individual variability. Epigenetic phenomena reshape gene function based on exposures and developmental or stochastic phenomena and should be monitored as well as indicators of a health status change.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 1

Recent Examples of Genetically-Guided Diagnoses for Rare and Idiopathic Conditions (‘Diagnostic Odysseys’) as well as Published Individual Monitoring Studies for Detecting Early Signs of Disease (‘Intensive Monitoring’)

Category	Source	Subject	Significance	Purpose of study	Variables/Endpoints	Results
Diagnostic Odyssey	Worthey EA et al. 2011 (75)	15 month old boy with intestinal illness similar to Crohn's disease	The first successful use of whole exome sequencing to identify a disease-causing genetic mutation.	Whole exome sequencing of a single subject with refractory inflammatory bowel disease.	Identification of the specific mutation causing severe intestinal inflammation.	A missense mutation was identified in a gene that played a role in the inflammatory response. A stem cell transplant cured the subject of the disease.
	Bainbridge MN et al. 2011 (76)	14 year old twins with a complex movement disorder	Further evidence of the benefits of whole genome sequencing in treating complex disease.	Whole genome sequencing of twins with a complex movement disorder.	Identification of genetic mutations responsible for the debilitating neurologic disease.	A compound heterozygous mutation was identified in a gene responsible for dopamine and serotonin synthesis. Supplementation with a serotonin precursor resolved symptoms.
	O'Rawe JA et al. 2013 (77)	37 yo male with severe obsessive compulsive disorder (OCD).	Highlights the use of whole genome sequencing in neuropsychiatric disease.	Whole genome sequencing of a single subject with refractory OCD.	Identification of genetic markers associated with neuropsychiatric illness.	Genetic variants were discovered in at least three genes associated with neuropsychiatric disease. No actionable variants identified.
	Chen YZ et al. 2014 (78)	2 children with childhood onset autosomal dominant familial dyskinesia with facial myokymia (FDFM)	Identification of a new missense mutation as a contributing factor to this debilitating complex movement disorder	Whole exome sequencing of two children with FDFM.	Identification of a likely causal mutation for the debilitating neurologic disease.	Both subjects had a similar mutation identified in the ADCY5 gene.
	Wartman LD 2015 (79)	20+ yo male with acute lymphoblastic leukemia (ALL)	One of the first instances of using whole genome sequencing to identify a drug target for ALL	Whole genome sequencing of a single subject with recurrent ALL.	-Identification of genetic mutations that could be used as targets for chemotherapy drugs. -Remission of acute lymphoblastic leukemia.	Subject did go into remission after a targetable mutation was identified. He was then able to receive second stem cell transplant.
Intensive Monitoring	Chen R et al. 2012 (80)	54 yo healthy male	First report of an integrative personal omics profile (iPOP) with genomic, transcriptomic, proteomic, metabolomic, and autoantibody profiles	Longitudinal monitoring of multiple omics profiles to assess health and disease states over a 14 month period	Multiple factors analyzed and correlated with subject's activity levels, diet, ingestion of medications, and development of infections or other disease	Disease risk could be assessed from genomic sequencing based on development of infections then Type 2 diabetes. Dynamic changes noted in omics that could be useful for early disease detection and prevention
	Smarr, L. 2012 (81)	60+ yo male with intestinal disease	Comprehensive monitoring for a full decade, highlighting a systems biology approach.	Quantified self approach to disease detection and tracking of treatment.	Genomics plus longitudinal monitoring of standard biomarkers, lifestyle, and microbiome assessment.	Time series measurements of inflammatory markers and stool microbiome led to diagnosis of subject's Crohn's Disease, thus proving clinical utility of the quantified self approach.
	David LA et al. 2014 (82)	Two healthy males, ages unknown	One year of daily gut and saliva microbiome analysis	Investigation of how lifestyle can affect an	Diet, exercise and travel data were measured, in addition to	Microbial communities were stable overall except during

Category	Source	Subject	Significance	Purpose of study	Variables/Endpoints	Results
				individual's microbial communities.	the daily evaluation of gut and saliva microbial communities.	periods of enteric infection (food poisoning) and travel out of the United States, when profound changes were observed.
	Forsdyke DR 2015 (83)	60 year old male with hypertension	Discovery of seasonal changes in sensitivity to antihypertensives in a single individual.	Determine if blood pressure changes in the summer time would be significant enough to warrant adjusting dosages of various antihypertensive medication.	Daily or twice daily blood pressure measurements were correlated with outdoor temperatures for over 12 years.	Summertime sensitivity to the angiotensin II receptor blocker losartan was identified requiring dosage adjustment to prevent side effects of hypotension.
	Trammell SA et al. 2016 (84)	52 yo healthy male	First clinical trial of the pharmacokinetics of nicotinamide riboside (NR) in humans	Determine bioavailability of NR when taken as a precursor vitamin of nicotinamide adenine dinucleotide (NAD+)	Subject took daily doses of NR then had multiple measurements of blood levels of NAD+ and its metabolites.	NAD+ levels did rise after ingestion NR and a new metabolite, nicotinic acid adenine dinucleotide (NAAD), was identified in the NR to NAD+ pathway.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/337696975

A road map for understanding molecular and genetic determinants of osteoporosis

Article *in* Nature Reviews Endocrinology · December 2019

DOI: 10.1038/s41574-019-0282-7

CITATIONS
316

READS
584

8 authors, including:

Tie-Lin Yang
Xi'an Jiaotong University
192 PUBLICATIONS **6,534** CITATIONS
SEE PROFILE

Hui Shen
Tulane University
655 PUBLICATIONS **95,560** CITATIONS
SEE PROFILE

Anqi Liu
Tulane University
40 PUBLICATIONS **485** CITATIONS
SEE PROFILE

Shan-Shan Dong
Xi'an Jiaotong University
134 PUBLICATIONS **5,735** CITATIONS
SEE PROFILE

All content following this page was uploaded by Shan-Shan Dong on 09 December 2019.

The user has requested enhancement of the downloaded file.

A road map for understanding molecular and genetic determinants of osteoporosis

*Tie-Lin Yang[1], Hui Shen[2], Anqi Liu[2], Shan-Shan Dong[1], Lei Zhang[3], Fei-Yan Deng[3], Qi Zhao[4] and Hong-Wen Deng[2,5]**

Abstract | Osteoporosis is a highly prevalent disorder characterized by low bone mineral density and an increased risk of fracture, termed osteoporotic fracture. Notably, bone mineral density, osteoporosis and osteoporotic fracture are highly heritable; however, determining the genetic architecture, and especially the underlying genomic and molecular mechanisms, of osteoporosis in vivo in humans is still challenging. In addition to susceptibility loci identified in genome-wide association studies, advances in various omics technologies, including genomics, transcriptomics, epigenomics, proteomics and metabolomics, have all been applied to dissect the pathogenesis of osteoporosis. However, each technology individually cannot capture the entire view of the disease pathology and thus fails to comprehensively identify the underlying pathological molecular mechanisms, especially the regulatory and signalling mechanisms. A change to the status quo calls for integrative multi-omics and inter-omics analyses with approaches in 'systems genetics and genomics'. In this Review, we highlight findings from genome-wide association studies and studies using various omics technologies individually to identify mechanisms of osteoporosis. Furthermore, we summarize current studies of data integration to understand, diagnose and inform the treatment of osteoporosis. The integration of multiple technologies will provide a road map to illuminate the complex pathogenesis of osteoporosis, especially from molecular functional aspects, in vivo in humans.

**e-mail: hdeng2@tulane.edu*

https://doi.org/10.1038/s41574-019-0282-7

Osteoporosis, the most common bone disorder worldwide (FIG. 1), is characterized by low bone mineral density (BMD) and an increased risk of osteoporotic fracture[1]. According to the WHO, osteoporosis is defined as a BMD that lies 2.5 standard deviations or more below the average value for young healthy women (T-score $\leq$ 2.5)[2]. Consequently, the clinical diagnosis and assessment of osteoporosis is mainly based on measurements of BMD[3]. Of note, BMD has a heritability of 0.6–0.8, meaning that 60–80% of the variation in BMD is inherited from parents and the remainder is derived from the environment[4]. In addition, osteoporotic fracture, which is the end point clinical outcome of osteoporosis, has a heritability of 0.5–0.7 (REF.[5]). Despite this strong heritability, determining the genetic architecture (BOX 1), and especially the underlying genomic and molecular mechanisms of osteoporosis in vivo in humans, is challenging.

Genome-wide association studies (GWAS) have identified hundreds of susceptibility loci for osteoporosis[6]. However, the causal variants and/or genes and their molecular functional mechanisms are largely unknown. Other omics technologies, that is, transcriptomics, epigenomics, proteomics, metabolomics and metagenomics (BOX 1), individually might provide a useful glimpse into the windows of osteoporosis pathophysiology. However, data generated by single-omics technologies are not comprehensive enough to capture the complete pictures of the molecular events leading to osteoporosis. A change in the status quo calls for integrative multi-omics and/or inter-omics analyses with approaches in systems genetics and genomics.

Compared with individual-omics studies, multi-omics studies can provide pathways of information from the original root cause of a disease (for example, genetic variations) to functional consequences or relevant interactions[7,8]. Therefore, multi-omics data can provide a clearer and more comprehensive view of the pathogenesis of osteoporosis. This knowledge is important as it is not currently possible to accurately identify all patients who will experience osteoporotic fracture from measurements of BMD.

Key points

- Osteoporosis, which is the most common bone disorder worldwide, and its related traits (low bone mineral density and osteoporotic fracture) are highly heritable.
- Multiple omics technologies, including genomics, transcriptomics, epigenomics, proteomics and metabolomics, have been applied to identify the molecular factors contributing to the pathogenesis of osteoporosis.
- Building upon the success in single-omics discovery research, studies have integrated data from different omics levels to better elucidate the molecular and functional mechanisms for osteoporosis.
- Integration of omics approaches can provide a holistic road map to comprehensively illuminate the complex pathogenesis of osteoporosis and fulfil the potential of personalized disease risk prediction, intervention and treatment as well as drug development or re-purposing.

In this Review, we briefly introduce the mechanisms of bone homeostasis before providing a broad survey of GWAS and the multiple-omics studies on osteoporosis (FIG. 2). We discuss the advantages and challenges of each technology and then focus on studies in which multiple levels of omics are integrated to elucidate the potential molecular mechanisms for osteoporosis. We also discuss the current methods and challenges in combining and interpreting multi-omics data in osteoporosis studies.

Bone homeostasis

Bone homeostasis is mainly controlled by the action of osteoblasts, osteocytes and osteoclasts; bone is a highly metabolically active tissue, which undergoes a continuous cycle of bone formation mediated by osteoblasts and bone resorption facilitated by osteoclasts. Importantly, disruption of bone homeostasis has a fundamental role in the pathogenesis of osteoporosis[9].

The cells that comprise bone tissue have diverse origins. For example, osteoblasts are derived from mesenchymal stem cells, which can also give rise to adipocytes, chondrocytes and myocytes[10] (FIG. 3). Osteoblasts produce bone by synthesizing extracellular matrix consisting of various proteins, the most abundant being type I collagen. The extracellular matrix is known as the osteoid when first deposited and is subsequently mineralized through the accumulation of calcium phosphate as hydroxyapatite ($Ca_{10}(PO_4)_6(OH)_2$). Multiple transcription factors (such as RUNX2 and OSX) and the major developmental signals (such as WNT signalling) are reported to regulate osteoblast differentiation and function[11]. By contrast, the bone resorbing osteoclasts are large, multinucleated cells formed by the fusion of precursors from the monocyte-derived macrophage lineage[12] (FIG. 3). Osteoclasts can dissolve minerals and digest bone matrix by secreting hydrochloric acid and proteolytic enzymes[13]. Similar to osteoblasts, multiple factors are reported to regulate osteoclast differentiation and/or function, including macrophage colony-stimulating factor (M-CSF), receptor activator of nuclear factor-κB ligand (RANKL), cytokines (for example, IL-1) and αVβ3 integrin[14].

Author addresses

[1]Key Laboratory of Biomedical Information Engineering of Ministry of Education, School of Life Science and Technology, Xi'an Jiaotong University, Xi'an, China.
[2]Center of Bioinformatics and Genomics, Department of Global Biostatistics and Data Science, Tulane University, New Orleans, LA, USA.
[3]Center for Genetic Epidemiology and Genomics, School of Public Health, Medical College of Soochow University, Jiangsu, China.
[4]Department of Preventive Medicine, College of Medicine, University of Tennessee Health Science Center, Memphis, TN, USA.
[5]School of Basic Medical Science, Central South University, Changsha, China.

The major cellular component of bone tissue is osteocytes, which are cells that originate from osteoblasts; these cells are deeply embedded in bone tissue and comprise >90% of all bone cells[15]. Osteocytes can orchestrate bone homeostasis by regulating the function of bone-forming osteoblasts and bone-resorbing osteoclasts. For example, osteocytes regulate bone formation by secreting modulators of the WNT signalling pathway such as activators (nitric oxide and ATP) and inhibitors (sclerostin and DKK1)[16]. Moreover, osteocytes also express RANKL and M-CSF to increase osteoclast activity as well as osteoprotegerin and nitric oxide to inhibit osteoclast formation and activity[16]. A detailed description of the crosstalk between the three major bone cellular subsets falls outside the scope of this Review, but can be found in REF.[17].

The genome influences the maintenance of bone homeostasis by encoding many factors that modulate the differentiation and activities of bone cells. In addition to several components in the aforementioned signalling pathways, hundreds of common genetic variants associated with BMD or osteoporotic fracture have been discovered. Beginning in the next section, we will present an update on what is known about genetic risk factors for osteoporosis.

Susceptibility loci derived from GWAS

Mainly through testing millions of single nucleotide polymorphisms (SNPs), GWAS have been used to identify over 500 susceptibility loci for osteoporosis-related traits[18–20] (Supplementary Table 1). The sample sizes of early single-sample GWAS were fairly small (<10,000)[21–24], resulting in limited statistical power; indeed, only risk loci with an effect size (BOX 1) ranging from 7.8×10^{-4} to 1.4×10^{-2} could be identified[21–24] (BOX 2). However, with the increases in sample size and improvements in statistical power, GWAS meta-analyses can identify more risk loci with effect sizes of 10^{-4} to 10^{-3} (BOX 2). For example, the GEnetic Factors for OSteoporosis (GEFOS) consortium, which consists of unrelated participants and sample sizes ranging from 19,195 to 66,628, has identified dozens of new susceptibility loci (such as *EN1*)[25] and has replicated some risk loci identified by earlier GWAS[26–30] (such as *ESR1* and *LRP5*). In mice, *En1* expression was only observed in osteogenic lineages, and conditional loss of *En1* resulted in low bone mass[25]. As the oestrogen receptor, ESR1 affects bone formation in both osteoblasts progenitors[31] and mature osteoblasts[32]. ESR1 could also induce the transcription of Fas ligand in osteoblasts, resulting in a paracrine signal to induce osteoclast apoptosis[31]. In the canonical WNT signalling pathway in bone formation, the activation of β-catenin through LRP5 and Frizzled results in the upregulation of transcription factors that are crucial for osteoblast differentiation[33] (FIG. 3). Together, the loci identified from GEFOS cumulatively explained ~6% of the phenotypic variance in BMD.

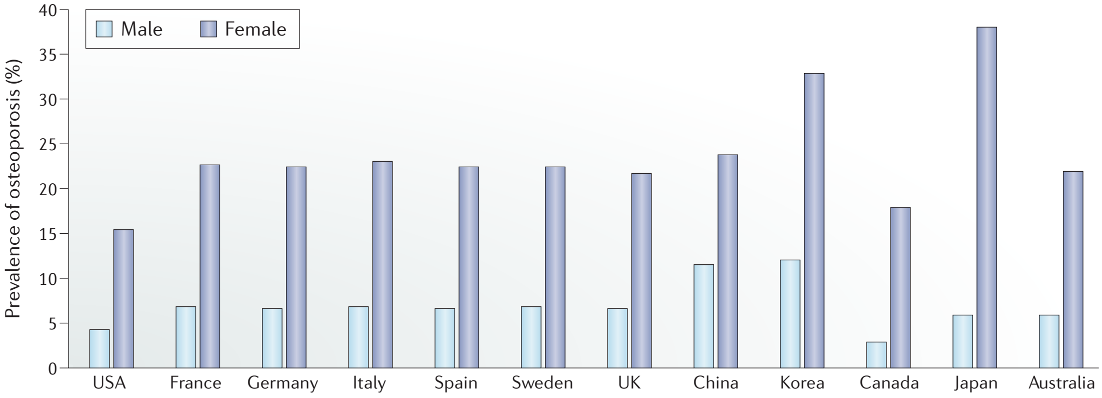


Fig. 1 | **Prevalence of osteoporosis in populations of age 50 years and older in selected countries.** The prevalence of osteoporosis in the non-institutionalized USA population was calculated using data collected by the National Health and Nutrition Examination Survey 2005–2010 (REF.[153]). The statistics for six European countries (France, Germany, Italy, Spain, Sweden and the UK) were retrieved from a report by the International Osteoporosis Foundation[154]. The statistics for China and Korea were obtained from a meta-analysis study published in 2016 (REF.[155]) and the Korea National Health and Nutrition Examination Survey 2008–2010 (REF.[156]), respectively. Data for Canada, Japan and Australia were obtained from a 2014 study[157].

The UK Biobank project[34] is remarkable due to its huge number of samples obtained from participants from the UK general population. Importantly, two GWAS in ~140,000 and 420,000 UK Biobank participants have identified 203 and 518 loci, respectively, associated with variation in ultrasound-based measurements of estimated heel BMD (eBMD)[6,35] (FIG. 3). For example, DAAM2 was identified in a UK Biobank study; reduction of DAAM2 protein levels resulted in reduced bone mineralization[6]. In addition, the largest UK Biobank cohort has nearly 100% power to identify loci with an effect size of 1×10^{-5} (BOX 2). The loci identified from studying the UK Biobank cohort can cumulatively explain 20% of the phenotypic variance in eBMD, therefore substantially increasing our understanding of the genetic architecture of osteoporosis.

Several groups have also used GWAS to carry out age-specific and sex-specific analyses[30,35]. Variants in *ESR1* and in close proximity to *RANKL* showed a clear age-dependent effect on BMD[30]. A single variant, rs17307280, at *FAM9B* on the X chromosome was significantly associated with eBMD in men only[35]. The function of FAM9B in bone homeostasis is still unknown. Of note, a direct link exists between oestrogen deficiency after menopause and the development of osteoporosis[36] in women. Consequently, researchers have also attempted to investigate specific risk loci in premenopausal[37] and postmenopausal[38] women. SNPs in *WNT16* and *ESR1* were identified to be associated with lumbar spine BMD in premenopausal women[37]. *WNT16* is a member of the WNT signalling pathway and *Wnt16* knockout mice have reduced bone strength[39]. In postmenopausal women[38], variants in *TNFRSF11B*, *SPTBN1*, *ESR1* and *LRP4* were reported to be susceptibility loci for osteoporotic fracture. The protein encoded by *TNFRSF11B* is osteoprotegerin, which functions as a decoy receptor of RANKL to inhibit bone resorption[40]. In vivo evidence suggested that LRP4 in osteoblasts could suppress bone formation and promote osteoclastogenesis through acting as a receptor of sclerostin[41].

The missing heritability problem. The BMD-associated loci identified so far in GWAS still cannot account for all the heritability in osteoporosis. Consideration of gene–environment interactions might explain part of this ‘missing heritability’[42]. Researchers could explore gene–environment interactions through investigating the effects of genetic variants on osteoporosis in specific population subgroups affected by specific environmental factors (for example, smoking). In addition, many alternative approaches have been developed to serve as complementary methods for the standard SNP-based GWAS described earlier in this Review. For example, copy number variation-based GWAS focus on identifying variation in gene copy number. The first copy number variation-based GWAS identified that a deletion variant of *UGT2B17* was associated with hip osteoporotic fracture[43]. *UGT2B17* encodes an enzyme catabolizing steroid hormones that have effects on bone formation[43].

Another useful approach is that used in gene-based GWAS, which regard a gene as a basic unit for association analyses, thereby improving the statistical power (BOX 2) by overcoming the potential allelic heterogeneity (BOX 1) and reducing the number of independent statistical tests from millions (for SNPs) to tens of thousands (for genes)[44]. A 2018 study found three novel BMD-associated genes (*UBTF*, *AAAS* and *C11orf58*) through a gene-based GWAS strategy[44]. The roles of the proteins encoded by these three genes in bone homeostasis are still unknown. A powerful technique for identifying biological pathways that might be involved in a heritable trait is pathway-based GWAS, which considers SNPs in multiple genes in a related pathway jointly. The first pathway-based GWAS for osteoporosis identified a pathway that regulates autophagy as being associated with ultradistal radius BMD[45]. In addition to overcoming the effect of allelic heterogeneity and improving statistical power, gene-based GWAS and/or pathway-based GWAS also enable direct comparison between different populations with different linkage disequilibrium patterns and/or functional alleles[46].

Box 1 | **Key terms in genetic and omics studies**

Allelic heterogeneity
Multiple single nucleotide polymorphisms within the same gene and/or pathway jointly affect the same trait.

Distant gene
If a genetic variant affects the expression or otherwise interacts with genes other than the nearest gene, the target genes are referred as distant genes of the variant of interest.

Effect size
The portion of phenotypic variance that is explained by the tested variant.

Epigenomics
The study of genome-wide reversible modifications of DNA or DNA-associated proteins such as DNA methylation, histone acetylation and chromatin organization.

Expression quantitative trait loci (eQTL) analysis
A technique for assessing the associations between transcript expression and genotype to identify genetic variants that explain the variation in gene expression levels.

Fingerprint
Specific expression profiles of proteins, which can be used as characteristics to distinguish different individuals.

Genetic architecture
The characteristics of genetic variation that are responsible for heritable phenotypic variability[150].

Genome-wide association studies (GWAS)
Studies using a hypothesis-free method to investigate the associations between genetic variants and traits, including diseases.

Hybrid mouse diversity panel
A collection of approximately 100 well-characterized inbred strains of mice that can be used to analyse the genetic and environmental factors underlying complex traits.

Knowledgebase
A library used to store complex structured and unstructured information by a computer system.

Long-range
The distance between regulatory regions and their target genes is considered far, usually >100 kb.

Mendelian randomization
Mendelian randomization is a method of using genetic variants to determine whether an observational association between a risk factor and an outcome is consistent with a causal effect.

Metabolomics
A field of omics science to systematically measure small molecules, commonly knowns as metabolites, within cells, biofluids, tissues or organisms.

Ome
The objects of one field of study in biology, for example, the genome, proteome or metabolome.

Proteomics
The identification and quantification of the entire protein complement of a cell, tissue or organism under a specific, defined set of conditions.

Transcriptomics
The study of the complete set of RNA transcripts (including messenger, transfer, ribosomal and non-coding regulatory RNAs) produced by a cell or tissue under specific conditions.

Weighted gene co-expression network analysis (WGCNA)
A commonly used systems biology method for studying the correlation patterns among genes. WGCNA can be used for finding clusters (referred to as modules) of genes sharing similar expression patterns across a set of samples, and for relating modules to disease status and sample traits using module behaviour and topology characteristics.

Bivariate GWAS aim to discover SNPs associated with two different traits and could identify pleiotropic genes as well as the key mechanistic links between two diseases. Any identified pleiotropic genes could provide insight into a genetic mechanism shared by both traits as well as potentially facilitating repurposing of conventional drugs used for one disease for new use in the other. A relevant example in osteoporosis research was a study identifying SNPs in *SOX6* (REF.[47]) and the *TOM1L2–SREBF1* locus[48], which have pleiotropic effects on both BMD and lean mass. Finally, summary databased approaches, such as joint association studies[49], pleiotropic conditional false discovery rate method[50] and co-expression network prediction[51], which are methods that use GWAS summary statistics alone, have also identified a large amount of novel BMD loci.

Clinical translation of GWAS findings. To address the missing heritability problem, researchers could explore more loci under an omnigenic model[52], which proposes that gene regulatory networks are so interconnected that most heritability can be explained by effects on genes outside core pathways. New findings will provide novel insights into the genetic mechanism of osteoporosis as well as osteoporosis risk prediction. For example, two studies[53,54] have reported that genetic profiling of BMD-associated SNPs could improve the accuracy of osteoporotic fracture prediction. In addition, a genetic algorithm that uses SNPs as predictors also showed a strong correlation with eBMD (correlation coefficient $R = 0.415$)[55]. By contrast, for clinical applications such as drug target design, loci with large genetic effects on BMD and osteoporotic fracture risk might be more valuable. We should spare no effort to dissect the clinical applications of known risk loci, which is an ultimate goal of all GWAS.

The success of GWAS demonstrates the usefulness of disease genomics to translational medicine. Generally, compared with those without genetic support, drug targets that are reported as susceptibility loci in GWAS are twice as likely to succeed in clinical trials[56]. For osteoporosis, GWAS have provided genetic support for five of the eight (63%) approved anti-osteoporosis therapeutics. For example, *RANKL* is the drug target of denosumab and the genetic association between *RANKL* and BMD has been reported by GWAS[18]. In addition, GWAS results could provide resources for drug repositioning. Some of the susceptibility loci for BMD and osteoporotic fracture are shared with many other diseases and, consequently, some osteoporosis-associated genes have already been used as drug targets for other diseases (Supplementary Table 2). For example, the eBMD-associated gene *ACHE* is an approved drug (donepezil) target for Alzheimer disease. Accumulating evidence[57,58] has indicated that osteoporosis and hip fracture are common complications observed in patients with Alzheimer disease, although the mechanisms underlying this association remain poorly understood. In addition, as an inhibitor of ACHE, donepezil might serve as a potential drug for osteoporosis as it could prevent RANK-induced bone loss via inhibition of osteoclast differentiation[59].

The challenges of GWAS. Several challenges exist for the use of GWAS in osteoporosis, which should be addressed in future studies. For example, diverse phenotypes (BMD at different skeletal sites using different measurements) are now included in osteoporosis GWAS, resulting in inconsistent results. The inconsistencies might be due to the heterogeneity among the different types of measurements of BMD, as the genetic correlation coefficients between the commonly used dual-energy X-ray absorptiometry and other measurements (for example, ultrasonography and quantitative CT) range from 0.505 to 0.917 (REF.[60]). However, even when studies use the same type of measurement, there is still heterogeneity between BMD measured at different skeletal sites. In addition, as the ultimate goal of osteoporosis research is to reduce the incidence and prevalence of osteoporotic fracture, GWAS should use osteoporotic fracture as a direct phenotype. Several novel genes (such as *FAM210A*, *GRB10* and *ETS2*) have been identified as susceptibility loci by GWAS directly comparing individuals who have had an osteoporotic fracture and healthy controls[61]. The specific functions of these genes in bone homeostasis are still unknown. However, most of these studies combined heterogeneous forms of osteoporotic fracture at different skeletal sites, which have different aetiologies; the results obtained might therefore entail both false positive or false negative findings, which are difficult to discern. Two studies examined osteoporotic fracture at specific sites: vertebral fracture[62] and hip fracture[63]. Two SNPs (rs2468531 and rs12742784) located in the non-coding region were found to be associated with vertebral fracture[62]. One SNP (rs13182402) within the *ALDH7A1* gene was associated with hip fracture. The ALDH7A1 protein degrades and detoxifies acetaldehyde, which inhibits osteoblast proliferation and results in decreased bone formation[63]. We suggest that vertebral fracture might not be the best phenotype to study, as no consensus definition of vertebral fracture currently exists. By contrast, hip fracture has a consensus definition and might represent a feasible alternative for study.

The replication of findings in distinct GWAS is a routine process for validating the statistical results from the discovery sample[64,65]. However, inconsistent findings that occur between different studies might be caused by limited statistical power (BOX 2) or heterogeneous effects[66] between distinct populations and ethnicities[64]. These effects could explain discrepancies between studies, and even the different findings between GEFOS-1 and GEFOS-2 (REFS[26,28]), which were conducted by largely the same research group. Future meta-analyses of GWAS should enable researchers to fully consider the heterogeneity.

Another potential issue of GWAS is the difficulty in determining the causal variants of a specific phenotype owing to complex patterns of linkage disequilibrium. Some statistical inference methods, such as PAINTOR[67], GCTA-COJO[68] and FINEMAP[69], have been applied for the fine-mapping of plausibly causal variants for BMD with high confidence[6,35,70]. However, most fine-mapping algorithms need reference genotype data to estimate the linkage disequilibrium relationship between genetic variants, which introduces inescapable bias when the reference data and original GWAS data are not perfectly matched[71].

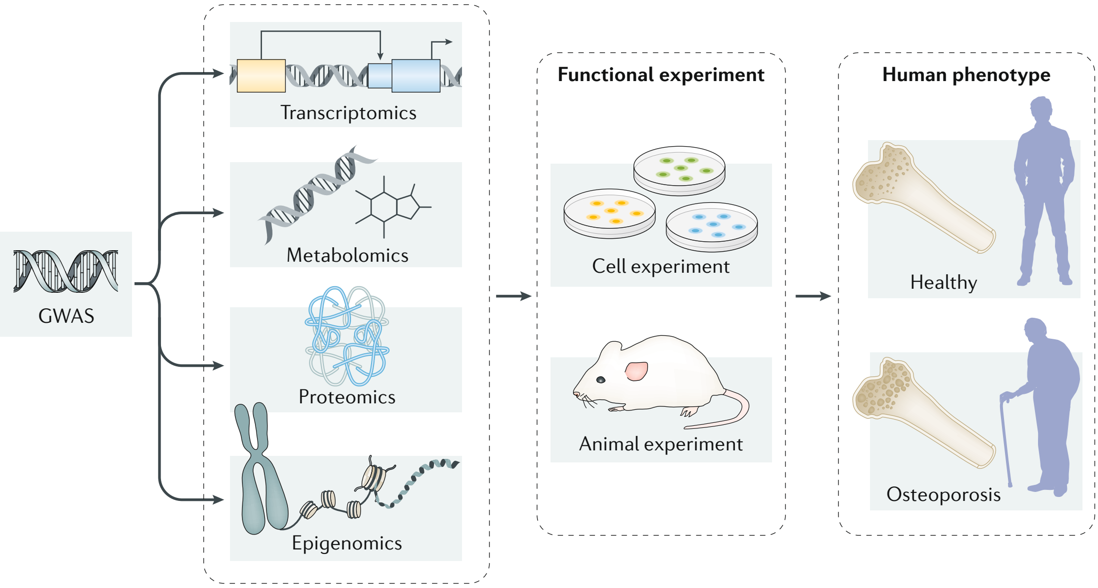


Fig. 2 | **Integrating multi-omics data to elucidate the molecular mechanisms of osteoporosis.** Multiple omics technologies, including genomics (mainly refers to genome-wide association studies (GWAS)), transcriptomics, epigenomics, proteomics and metabolomics, have been applied to dissect the pathogenesis of osteoporosis. Each technology individually can only provide limited insights into the biological mechanisms of osteoporosis. By integrating multiple omics data and following-up functional experiments in cell lines and/or animal models, researchers could capture a comprehensive view of the pathogenesis of this disorder.

Finally, although GWAS have identified hundreds of independent loci for osteoporosis-related traits (Supplementary Table 1), the underlying causal functional variants and biological regulatory mechanisms are still largely unknown. The use of GWAS alone might not be able to address these issues; however, considering functional genomics could provide new insights.

Insights from 'omics' studies

Functional genomics and other omics studies, which focus on, for example, global gene transcription, translation and protein–protein interactions, have identified a list of biomarkers for osteoporosis, providing insights into the biological pathways involved. Here, we review these studies, including transcriptomics, epigenomics, proteomics and metabolomics (BOX 1).

Transcriptomics. Transcriptomics studies are used to identify differentially expressed genes between patients with osteoporosis and control individuals. Further functional experiments are used to validate the roles of these genes in osteoblastic and osteoclastic differentiation and/or activity. So far, most of the research has studied the precursor cells of osteoblasts (that is,

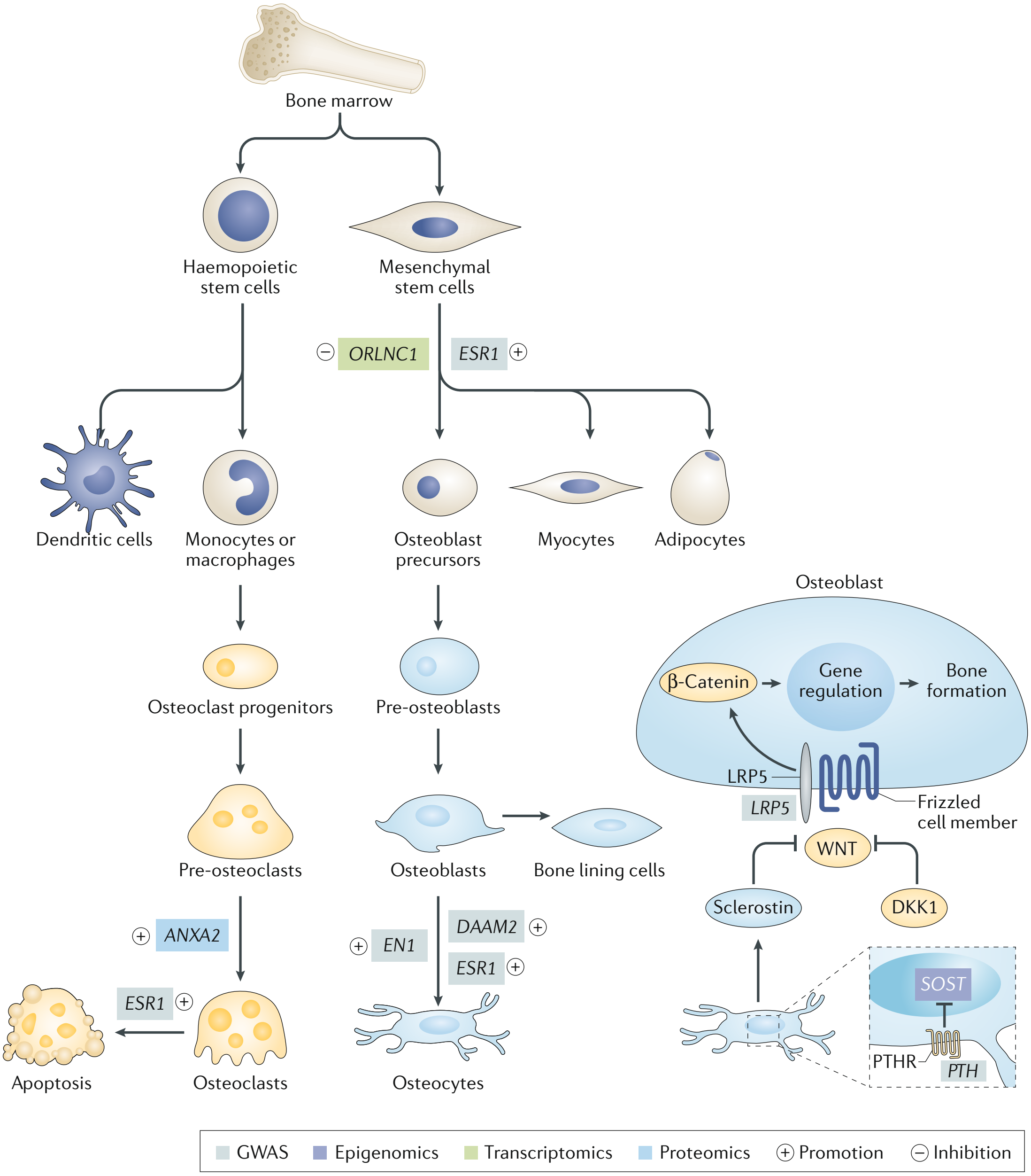

Bone marrow
Haemopoietic stem cells
Mesenchymal stem cells
⊖ ORLNC1
ESR1 ⊕
Dendritic cells
Monocytes or macrophages
Osteoblast precursors
Myocytes
Adipocytes
Osteoclast progenitors
Pre-osteoblasts
Pre-osteoclasts
Osteoblasts
Bone lining cells
⊕ ANXA2
ESR1 ⊕
⊕ EN1
DAAM2 ⊕
ESR1 ⊕
Apoptosis
Osteoclasts
Osteocytes
Osteoblast
β-Catenin
Gene regulation
Bone formation
LRP5
LRP5
Frizzled cell member
WNT
Sclerostin
DKK1
SOST
PTHR
PTH
GWAS
Epigenomics
Transcriptomics
Proteomics
⊕ Promotion
⊖ Inhibition

Fig. 3 | **Differentiation process of osteoblasts and osteoclasts.** Bone is a highly metabolically active tissue, which undergoes a continuous cycle of bone formation mediated by osteoblasts and bone resorption facilitated by osteoclasts. The cells that comprise bone tissue have diverse origins. Osteoblasts are derived from mesenchymal stem cells, which can also give rise to adipocytes, chondrocytes and myocytes. Osteoclasts are large, multinucleated cells formed by fusion of precursors derived from the monocyte–macrophage lineage. As the major cellular component of bone tissue, osteocytes originate from osteoblasts. We list several representative genes linked to bone metabolism by omics studies; functional experiments support their involvement in bone homeostasis. GWAS, genome-wide association studies; PTHR, parathyroid hormone receptor.

Box 2 | Statistical power calculations

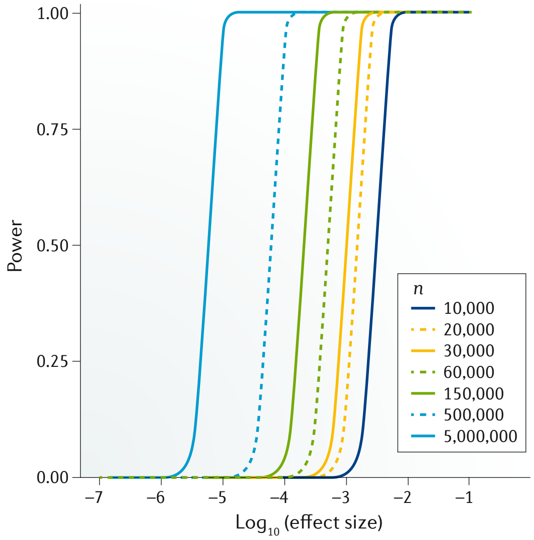
Statistical power is important in genome-wide association studies (GWAS) as it is the parameter that determines how effectively a given study will identify susceptibility loci. The statistical power of statistical significance is the probability that an association test statistic will reject a false null hypothesis. As an association test statistic asymptotically (that is, approaching a curve arbitrarily closely) follows a 1-degree of freedom, non-central χ^2 distribution, statistical power is solely determined by the non-centrality parameter λ of this distribution ($\lambda \geq 0$).

Statistical power is confounded by a variety of factors, including cryptic relatedness (individuals are more closely related to one another than assumed by the investigator) and population stratification[151]. In an ideal scenario in which an unrelated sample (that is, of individuals that are independent from each other) is drawn from a homogeneous population and under an additive mode of inheritance, the non-centrality parameter λ has a simple form[152]:

$$\lambda \approx \mathrm{N} \times \mathrm{e}$$

where N is the sample size and e is the effect size, which is defined as the portion of phenotypic variance that is explained by the target variant, that is, variant-level heritability.

Of note, statistical power is non-relevant to variant allele frequency. Therefore, to detect associated variants of $\frac{1}{n}$ of the original effect size, the sample size should increase to n times the original sample size to reach the same level of statistical power. Refer to the figure to see the power calculated under several typical sample size settings for a GWAS scan (type I error rate threshold $\alpha = 5 \times 10^{-8}$). For any given sample size, the range of effect size with intermediate power falls into a relatively narrow interval. For example, under the largest sample size shown in the figure (5,000,000), this GWAS setting has nearly 100% power to detect variants with effect size $> 1 \times 10^{-5}$ but nearly 0% power to detect variants with effect size $< 1 \times 10^{-6}$.



human bone marrow mesenchymal stem cells (BMSCs)) or osteoclasts (for example, circulating monocytes) (Supplementary Table 3).

The first comparative gene expression study of circulating monocytes was carried out in individuals with high or low lumbar spine BMD measured with dual-energy X-ray absorptiometry and identified three upregulated genes (*CCR3*, *HDC* and *GCR*) in participants with low BMD[72]. The products of these genes are CC-chemokine receptor 3 (CCR3), histidine decarboxylase and glucocorticoid receptor, respectively. They are all known to be involved in osteoclastogenesis[73–75]. In addition to mRNA, the levels of non-coding RNA, such as microRNA[76], long non-coding RNA (lncRNA)[77] and circular RNA[78], have also been studied. For example, a 2019 study[77] reported that lncRNA-ORLNC1 was overexpressed in bone tissues of patients with osteoporosis compared with controls, with a T-score ≥1, and were also upregulated in bone tissue, serum and BMSCs of ovariectomy-induced osteoporotic mice compared with the sham-operated group. Further functional studies in BMSCs of mice showed that a lncRNA-ORLNC1–miR-296–PTEN signalling axis functions as a critical regulator for the switch between the osteogenesis and adipogenesis of BMSCs, that is, whether or not these cells differentiate into osteoblasts or adipocytes (FIG. 3).

Except for a few studies focused on a particular gene, for example, a synonymous variant in exon 9 of *CD44* that might increase the susceptibility to osteoporosis by affecting the splicing mechanism[79], transcriptomics studies of osteoporosis generally do not consider alternative splicing. Of note, alternative splicing is a ubiquitous regulatory mechanism of gene expression and RNA mis-splicing is related to a large array of human diseases[80]. Therefore, transcriptomics studies that take alternative splicing into consideration are needed in the field of bone research.

Epigenomics. Epigenetic factors, mainly DNA methylation and histone modification, have considerable effects on the differentiation and activities of bone cells[81] and might contribute to pathogenetic mechanisms of osteoporosis[82]. The early epigenetic studies of osteoporosis often focused on a few candidate genes with known importance in bone biology (Supplementary Table 4). For example, one study compared the DNA methylation levels at the *SOST* promoter region in bone biopsy samples from postmenopausal women with osteoporosis ($n = 4$) and healthy controls ($n = 4$)[83]. They found that the *SOST* promoter showed higher DNA methylation in patients with osteoporosis than in controls, which was replicated in an independent cohort of 63 postmenopausal women (27 with osteoporosis and 36 without osteoporosis). Interestingly, the authors also observed strong positive associations between bone *SOST* mRNA and serum levels of sclerostin with BMD. Because sclerostin is known as an inhibitor of bone formation[84], the authors speculated that the increase in *SOST* promoter methylation and the reduction of *SOST* expression in patients with osteoporosis might represent a compensatory mechanism to counteracting osteoporosis-associated bone loss. However, conflicting results were reported by a 2019 study comparing femoral bone tissues from 16 Chinese patients with osteoporosis who had femoral neck and/or trochanter fractures and 16 controls (patients with traumatic factures but normal BMD), wherein *SOST* gene expression (at both mRNA and protein levels) was significantly increased and *SOST* promoter was slightly hypomethylated in patients with osteoporosis[85]. Therefore, the association between *SOST* methylation and expression with osteoporosis warrants further investigation.

Importantly, a few groups have explored the epigenome-wide association study (EWAS) of DNA methylation with osteoporosis in humans (Supplementary Table 4). For example, in 2013, the first EWAS of human bone detected 241 differentially methylated CpG sites in femoral head trabecular bone specimens between 27 patients with osteoporotic hip fractures and 26 patients with hip osteoarthritis[86]. Bone samples from true control individuals were not available owing to the ethical difficulties in obtaining bone biopsy samples from healthy people. So far, the largest EWAS using bone specimens were performed in 84 postmenopausal women with substantially varied BMD[87] and identified 63 differentially methylated CpGs associated with BMD.

Owing to the difficulty in obtaining human bone tissue samples, some EWAS (Supplementary Table 4) used whole blood as a proxy to test for the associations of DNA methylation with osteoporosis and BMD. One notable example study performed a large-scale EWAS for femoral neck and lumbar spine BMD in whole blood of 5,515 individuals[88]; however, the study did not reveal strong consistent association signals for DNA methylation at any of the >450,000 tested CpG sites. The lack of statistically significant associations between blood DNA methylation and osteoporosis was also reported in another independent EWAS[89]. Together, these findings suggest that blood DNA methylation patterns might not efficiently reflect the epigenomic status of bone cells if whole blood is used.

It is worth mentioning that aside from DNA methylation, other epigenetic features (for example, histone modification and high-order chromatin structure) have rarely been studied for osteoporosis[82]. Compared with many other human tissues and cell types, current data on the epigenetic architecture of human primary bone tissue and cells are rather scarce, even in large-scale consortium projects such as ENCODE[90] and NIH Roadmap[91].

Proteomics. To further investigate the molecular determinants of BMD, proteomics studies have been carried out to catalogue proteins expressed in bone-related cells or tissues, to identify proteins that are involved in the pathogenesis of osteoporosis, and to investigate protein biomarkers that are diagnostic of osteoporosis or predictive of osteoporotic fracture. Importantly, a 2008 study identified 38 differentially expressed BMD-associated proteins from the proteome of circulating monocytes in Chinese premenopausal women with discordant hip BMD[92]. The significance of ANXA2 and GSN proteins to osteoporosis was later cross-validated in independent cellular proteomics studies in white women with varying BMD[93,94]. Of note, ANXA2 could regulate osteoclastogenesis[93] (FIG. 3) and the function of osteoblasts[95], and increased plasma levels of this protein were shown to serve as a risk biomarker for osteoporosis and to predict osteoporotic fracture in Chinese individuals aged over 65 years[95]. The first comprehensive proteome knowledgebase (BOX 1) for human monocytes, developed in 2017 (REF.[96]), involves a total of 2,237 unique protein-encoding genes and provides a reference map for future in-depth research on monocyte biology and osteoporosis.

In addition to protein biomarkers developed from cellular proteomics studies, a number of potential protein biomarkers have been identified directly from plasma or serum proteomics studies. For example, one group identified a diagnostic fingerprint (BOX 1) for bone turnover in the serum of postmenopausal women, which could serve as a reflection of the increased osteoclast activity, leading to increased bone turnover, that is associated with decreasing BMD[97]. Besides, multiple layers of evidence were integrated to further ascertain the significance of proteins and genes of interest for osteoporosis. For example, a prior proteomics study discovered that superoxide dismutase 2 (SOD2) in circulating monocytes was significantly upregulated at protein level in vivo in Chinese individuals with low versus high hip BMD[92]. By integrating evidence from DNA, RNA and protein levels, a pursuant study ascertained *SOD2* as a susceptibility gene for osteoporosis in Chinese populations[98]. Furthermore, proteome-based network and pathway studies identified important groups of proteins for osteoporosis[99,100]. For example, a mass spectrometry-based quantitative proteomics study integrated with network analysis identified two novel pathways, that is, regulation of actin cytoskeleton and "leukocyte transendothelial migration" as being related to osteoporosis[96].

Some serum or plasma-derived protein biomarkers with diagnostic potential for osteoporotic fracture were identified from cross-sectional studies (Supplementary Table 5). Furthermore, serum-derived protein biomarkers predictive of incident osteoporotic fracture were identified from prospective studies (Supplementary Table 6). For example, a prospective study[101] suggested that increased serum levels of sclerostin are associated with increased risk of hip fracture. Excitingly, the antibody of sclerostin, that is, romosozumab (Evenity), has been approved by the FDA as a novel drug to treat osteoporosis in postmenopausal women with high risk of fracture.

Metabolomics. The human metabolome (BOX 1) reflects downstream changes in gene sequence and gene and/or protein expression. Thus, small changes in genes, their transcription and the translation of proteins that are hard to detect in other omics studies can be amplified at the metabolomic level[102], enabling the most powerful screening of biomarkers or therapeutic targets closely related to disease or phenotype. Of note, preclinical studies carried out in postmenopausal osteoporosis animal models have shown different metabolomic profiles compared with control animals, especially regarding changes in amino acid and lipid metabolism[103,104].

The first metabolomics study of osteoporosis in humans compared high and low BMD groups and reported four distinguishing metabolites: lactate, acetone, acetate and glutamine[105]. Subsequently, eight metabolomics studies investigating osteoporosis in humans have been published and have reported >100 metabolites associated with BMD (Supplementary Table 7). In line with the findings of animal studies, human studies also highlight the importance of amino acid and lipid metabolism in bone. In particular, many amino acids were reported to be associated with BMD by multiple independent studies[105–109]. For example, glutamine is increased[107,110] and proline is decreased in menopausal women with decreased BMD[109]. Moreover, carbohydrate[108] and nucleoside[106] metabolism are also altered in women with decreased BMD. Finally, BMD-related metabolic pathways have been reported, including bile acid biosynthesis, which might affect intestinal calcium absorption[107], and the tricarboxylic acid cycle, which might be related to increased oxidative stress in patients with low BMD[108].

Metabolomics research of osteoporosis is in the early stages. For example, all the current studies (Supplementary Table 7) are cross-sectional. Although two studies used Mendelian randomization (BOX 1) to provide potential evidence for a causal relationship of identified metabolites with osteoporosis[110,111], well designed longitudinal studies are necessary. In addition, potential sex differences in osteoporosis-related metabolites should be investigated, as most published studies only include women. For meaningful data generation, metabolite identity validation and absolute quantification should follow untargeted metabolomics analysis, which was used in all the current studies. These steps are necessary for conducting functional research of identified metabolites to derive meaningful biological knowledge[112].

Summary. Currently, functional genomics and other omics studies of osteoporosis have limitations, for example, the use of inappropriate tissue and/or cell specimens, small sample sizes and subsequent low statistical power, poorly characterized cohorts, and/or lack of healthy control individuals. Normal bone samples from healthy individuals are difficult to obtain as bone biopsy is invasive and carries a small risk of bleeding, fracture and/or infection. Consequently, bone samples obtained by biopsy from patients with osteoarthritis have been used as surrogate controls in several studies of osteoporosis[86]. However, the use of osteoarthritis bone samples might not be an ideal study design, as it is questionable to what extent cells from osteoarthritis bone resemble normal bone cells. Several studies have suggested that the trabecular structure and gene expression of osteoarthritis bone, at least in the vicinity of articular tissue, might differ from those of normal healthy bone[113,114].

In contrast to genetic variants, functional genomic and other omics changes are cell and tissue specific and therefore studies must be performed in bone or bone lineage cells. Moreover, bone itself is a heterogeneous tissue, composed of a number of different cell types (FIG. 3) and failure to account for the cellular heterogeneity might lead to false-positive or false-negative results[115,116]. A large consortium aiming to generate large-scale data for bone or bone lineage cells, such as the Cancer Genome Atlas Program for cancer, would greatly improve our ability to understand, diagnose, treat and prevent osteoporosis.

An integrative approach

Although each omics approach has provided useful information to begin to understand the pathophysiology of osteoporosis, no single omics approach (BOX 1) can capture the entire biological complexity of osteoporosis. Thus, approaches to integrate multi-omics data have emerged to create a holistic picture of osteoporosis. Next, we describe ways in which integrative multi-omics can help GWAS loci interpretation and the identification and prioritization of osteoporosis-related genes and pathways.

Integrating multi-omics data to interpret GWAS. The function of genetic variants located in coding regions of DNA, that is, in exons, is straightforward to study. However, the majority of GWAS susceptibility loci are located in non-coding or intergenic regions[117], which poses challenges to the functional interpretation of risk variants and the translation of GWAS findings into clinical application. Therefore, the first crucial step is to prioritize and/or pinpoint the true causal and/or functional variants[118].

A complementary approach to fine-mapping is to leverage multiple layers of epigenomic or functional data to assign potential functions for genetic variants identified by GWAS. For example, the assay for transposase-accessible chromatin using sequencing (known as ATAC-seq) is widely used to evaluate the accessibility of chromatin at the location of possible functional GWAS variants. Databases like RegulomeDB[119] can also be used to annotate SNPs with known and predicted regulatory elements such as regions of DNase I hypersensitivity and binding sites of transcription factors. The systematic analyses of the epigenetic features in all known osteoporosis risk loci identified by GWAS could identify common regulatory elements for osteoporosis-associated genes and might predict new osteoporosis-linked genes with similar regulatory features[120]. Using machine learning algorithms, researchers can extract intricate regulatory information underlying GWAS variants and use this knowledge to predict new osteoporosis susceptibility variants[121]. For example, one study predicted 37,584 new candidate osteoporosis-associated SNPs by applying a random forest model[121]. However, epigenetic regulatory analysis on GWAS variants have largely relied on reference epigenomic data (for example, NIH Roadmap Epigenomics Project) in which epigenomic profiles of each unique type of human tissues or cells were mostly generated from only one individual. Therefore, the allelic effects of GWAS variants on the inter-individual variation of cell-specific and/or tissue-specific epigenomic profiles (for example, individuals with different genotypes at a specific loci might have different levels of chromatin accessibility at the corresponding region in a specific tissue or cell of interest) are largely unknown. Future investigation with genomic and epigenomic data generated from the same set of a large number of participants should help uncover the allelic regulatory effects of osteoporosis-associated genetic variants on the epigenome.

Expression quantitative trait loci (eQTL) analysis (BOX 1) in osteoblasts[122] and osteoclasts[123] has helped to characterize the function of causal genetic variants located outside of exons and their downstream target genes in osteoporosis. For example, microarray profiles of undifferentiated osteoblasts from 95 unrelated donors of Swedish origin were used to identify eQTL for BMD GWAS[122]. Data in cultured primary osteoclasts from 158 individuals were used to identify colocalizing eQTL for genes associated with BMD[123]. Moreover, eQTL analysis on other bone-related cells and tissues, such as circulating monocytes[120], lymphocytes[124] or whole blood[124,125], have also been widely applied to decipher the regulatory roles of GWAS variants in osteoporosis. For example, lymphocytes and peripheral blood eQTL were used to identify *LINC00339* as a potentially causal gene for BMD GWAS[124]. The current sample sizes of eQTL studies on bone tissues are fairly small (<200); future studies in large samples are needed to increase the statistical power. In addition, the observed eQTL mapping on osteoporosis-associated GWAS variants might be accidental owing to high linkage disequilibrium with the true causal variant. To address this problem, several colocalization analysis approaches[126,127] have been developed to determine whether susceptibility loci identified by GWAS and eQTL share the same causal variant.

Importantly, studies in 3D chromatin interactions, which are measured by high-throughput chromatin interaction capture (known as Hi-C) analysis, have revealed a new regulatory mechanism by which functional GWAS variants could regulate distant gene expression via the formation of long-range (BOX 1) loops[6]. For example, the first Hi-C analysis in human osteoblasts and osteocytes revealed that the genes that had chromatin interactions with causal eBMD-associated variants were strongly enriched for osteocyte signature genes[6], suggesting that osteoporosis-associated genetic variants affect crucial bone-related genes by long-range regulation. A 2019 study further demonstrated that 57% of the identified chromatin interaction pairs linking osteoporosis-associated SNPs to gene promoters fell in distant genes only[128]. In one notable study, researchers found that five functional SNPs were located in an intergenic super-enhancer that could regulate the expression of *RANKL* (FIG. 3) via long-range chromosomal interactions (>100 kb)[129]. In addition, the intergenic non-coding osteoporosis-associated GWAS variant rs6426749 at 1p36.12 could regulate *LINC00339* expression via long-range loop formation (~ 360 kb)[124]. Downregulation of *LINC00339* significantly increases the expression of *CDC42* in osteoblasts, which is a pivotal regulator involved in bone metabolism[124].

When integrating multi-omics data to interpret osteoporosis GWAS susceptibility loci, the most commonly used method is to build evidence for a functional variant signal through the pairwise analyses of data sets[6,124]. Multi-dimensional methods that can analyse three or more data sets simultaneously (for example, Bayesian models[130]) will further advance our understanding of osteoporosis aetiology.

Gene and network-based studies. As genes do not function in isolation, the integration of different omics datasets at the level of genes and gene networks represents a critical approach to further increase our system-level understanding of the genetic basis of osteoporosis. For example, one group applied the weighted gene co-expression network analysis (WGCNA)[131] tool to transcriptomic expression data from bone tissues obtained from the Hybrid Mouse Diversity Panel[132] (BOX 1) and identified gene co-expression networks (termed as modules) that are highly associated with BMD[133]. From a specific BMD-associated module, they selected two 'hub' genes (that is, genes that are highly correlated with many other genes in a given module), *Maged1* and *Pard6g*, for subsequent in vitro and in vivo molecular function analyses. Specifically, knocking down the expression of *Maged1* and *Pard6g* in osteoblasts lead to altered osteoblast proliferation and differentiation, and mice deficient in *Maged1* had decreased BMD[133]. These results demonstrated the power of using a systems biology approach to discover novel genes and regulatory mechanisms involved in osteoporosis.

The network information obtained from the above analyses can be used to identify or prioritize novel disease-related genes that might not reach genome-wide significance in conventional GWAS analysis. For example, one group constructed a WGCNA network for 1,574 genes that had GWAS nominal evidence of association ($P < 0.05$) with BMD using gene expression data from transiliac bone biopsy samples obtained from 84 postmenopausal white women (39 with BMD T-score ≤ −1.0 and 45 with BMD T-score > −1.0) and discovered one module containing 44 highly interconnected genes significantly associated with BMD[134]. Further detailed submodule analyses suggested several novel candidate genes (for example, *HOMER1* and *SPTBN1*) that are important for bone metabolism[134]. *HOMER1* encodes a member of the Homer family of dendritic proteins and has an important role in glutamate signalling, which is involved in regulation of bone remodelling[135]. Interestingly, a 2017 study using a similar strategy also predicted the gene *SPTBN1* as a causal GWAS gene[51]. *SPTBN1* encodes a molecular scaffolding protein and is broadly expressed in many tissues and cell types, including high expression in osteoblasts[51]. Moreover, *Sptbn1* heterozygous knockout (*Sptbn1*$^{+/-}$) mice displayed significantly altered whole body BMD compared with the controls[51], confirming the critical role of this gene in the regulation of BMD. In addition to WGCNA, other more sophisticated algorithms for network construction and analysis of multi-omics data are quickly emerging and therefore provide exciting opportunities for future osteoporosis research[136,137].

Validation through functional studies

With numerous genes identified from GWAS, different omics studies and integrative analyses, the next step is to validate their functional roles in the pathogenesis of osteoporosis. The development of the CRISPR–Cas9 genome editing tool[138] has revolutionized molecular biology and genetics. For osteoporosis-associated gene functional studies, CRISPR–Cas9 has been used for coding-gene knockout[139], gene mutation repair[140] and validation of the effects of non-coding enhancers[124,129]. For example, CRISPR–Cas9-mediated knockouts of *DAAM2* in osteoblast cell lines resulted in a marked reduction in inducible mineralization[6].

Databases that record the consequences of disrupting each of the protein-coding genes in mice provide rich resources for validating the functional effects of candidate genes derived from GWAS and different omics studies. For example, The Origins of Bone and Cartilage Disease project, which is part of the International Knockout Mouse Consortium, aims to test over 1,600 mutant C57BL/6 strains with systemic single gene deletions for abnormalities of bone and cartilage. Although not specific to bone, similar databases, such as Mouse Genome Informatics and International Mouse Phenotyping Consortium, also provide integrated genetic, genomic and biological data to facilitate the study of osteoporosis. In addition to acting as validation resources, the identification of extreme skeletal phenotypes in mutant mouse lines in these databases can also serve as a new approach for osteoporosis-associated gene discovery[141]. For example, a 2014 study[142] analysed high-throughput screening data for 3,762 distinct global gene knockout mouse lines and identified multiple genes affecting bone mass.

Conclusions

As mentioned above, functional genomics studies of osteoporosis need to be conducted using skeletal samples and should also account for cellular heterogeneity, both of which pose obvious difficulties for human studies. The recent rapid development of single-cell omics technologies provides us with powerful new tools to individually or even simultaneously (that is, single-cell multi-omics approaches) assess various omics features at the resolution of individual cells[143], which can revolutionize our understanding of cellular heterogeneity. For example, using single-cell RNA sequencing, a 2018 study was able to successfully identify the long-sought multipotent human skeletal stem cell that exclusively differentiates into bone, cartilage and stroma but not adipose[144]. Application of the single-cell omics approaches to the field of osteoporosis research will enable us to explore known and unknown cellular heterogeneity in bone cells and tissues and dissect the cell-intrinsic and cell-specific mechanisms underlying osteoporosis.

To bridge the gap between the statistical associations and biological regulatory mechanisms, various experimental approaches can be used to test the function of osteoporosis susceptibility loci. Rather than testing the function of genetic variants one-by-one, high-throughput methods developed in 2017 (REF.[145]) (for example, massively parallel reporter assay and self-transcribing active regulatory region sequencing) can be used to examine thousands of variants in a single experiment. Applying such high-throughput methods into the post-GWAS studies of osteoporosis will lay functional evidence and mechanisms over the association results.

Another area for future osteoporosis research is the study of the human gut microbiome[146] — a number of studies in animal models[147] and humans[148] have provided compelling evidence for the importance of the gut microbiome in bone metabolism and health. With the strong interactions known to exist between the gut microbiome and the host genome[149], future host and microbiome multi-omics integration studies might lead to a major breakthrough in the prediction and therapeutic treatment of osteoporosis.

Published online: 02 December 2019

1. Kanis, J. A. Diagnosis of osteoporosis. *Osteoporos. Int.* **7**, S108–S116 (1997).
2. World Health Organization. Assessment of fracture risk and its application to screening for postmenopausal osteoporosis : report of a WHO study group [meeting held in Rome from 22 to 25 June 1992] (WHO, 1994).
3. Johnell, O. et al. Predictive value of BMD for hip and other fractures. *J. Bone Miner. Res.* **20**, 1185–1194 (2005).
4. Peacock, M., Turner, C. H., Econs, M. J. & Foroud, T. Genetics of osteoporosis. *Endocr. Rev.* **23**, 303–326 (2002).
5. Deng, H. W. et al. Relevance of the genes for bone mass variation to susceptibility to osteoporotic fractures and its implications to gene search for complex human diseases. *Genet. Epidemiol.* **22**, 12–25 (2002).
6. Morris, J. A. et al. An atlas of genetic influences on osteoporosis in humans and mice. *Nat. Genet.* **51**, 258–266 (2019). **As the largest GWAS of BMD until 2019, this paper performed analyses in over 426,000 individuals from the UK Biobank and identified 518 loci associated with BMD, 301 of which were novel**.
7. Sun, Y. V. & Hu, Y. J. Integrative analysis of multi-omics data for discovery and functional studies of complex human diseases. *Adv. Genet.* **93**, 147–190 (2016).
8. Hasin, Y., Seldin, M. & Lusis, A. Multi-omics approaches to disease. *Genome Biol.* **18**, 83 (2017).
9. Raisz, L. G. Pathogenesis of osteoporosis: concepts, conflicts, and prospects. *J. Clin. Invest.* **115**, 3318–3325 (2005).
10. Pittenger, M. F. et al. Multilineage potential of adult human mesenchymal stem cells. *Science* **284**, 143–147 (1999).
11. Long, F. Building strong bones: molecular regulation of the osteoblast lineage. *Nat. Rev. Mol. Cell Biol.* **13**, 27–38 (2011).
12. Udagawa, N. et al. Origin of osteoclasts: mature monocytes and macrophages are capable of differentiating into osteoclasts under a suitable microenvironment prepared by bone marrow-derived stromal cells. *Proc. Natl Acad. Sci. USA* **87**, 7260–7264 (1990).
13. Boyle, W. J., Simonet, W. S. & Lacey, D. L. Osteoclast differentiation and activation. *Nature* **423**, 337–342 (2003).
14. Teitelbaum, S. L. Bone resorption by osteoclasts. *Science* **289**, 1504–1508 (2000).
15. Knothe Tate, M. L., Adamson, J. R., Tami, A. E. & Bauer, T. W. The osteocyte. *Int. J. Biochem. Cell Biol.* **36**, 1–8 (2004).
16. Dallas, S. L., Prideaux, M. & Bonewald, L. F. The osteocyte: an endocrine cell and more. *Endocr. Rev.* **34**, 658–690 (2013).
17. Sims, N. A. & Walsh, N. C. Intercellular cross-talk among bone cells: new factors and pathways. *Curr. Osteoporos. Rep.* **10**, 109–117 (2012).
18. Sabik, O. L. & Farber, C. R. Using GWAS to identify novel therapeutic targets for osteoporosis. *Transl. Res.* **181**, 15–26 (2017).
19. Richards, J. B., Zheng, H. F. & Spector, T. D. Genetics of osteoporosis from genome-wide association studies: advances and challenges. *Nat. Rev. Genet.* **13**, 576–588 (2012).
20. Liu, Y. J., Zhang, L., Papasian, C. J. & Deng, H. W. Genome-wide association studies for osteoporosis: a 2013 update. *J. Bone Metab.* **21**, 99–116 (2014).
21. Styrkarsdottir, U. et al. Multiple genetic loci for bone mineral density and fractures. *N. Engl. J. Med.* **358**, 2355–2365 (2008).
22. Richards, J. B. et al. Bone mineral density, osteoporosis, and osteoporotic fractures: a genome-wide association study. *Lancet* **371**, 1505–1512 (2008).
23. Styrkarsdottir, U. et al. New sequence variants associated with bone mineral density. *Nat. Genet.* **41**, 15–17 (2009).
24. Xiong, D. H. et al. Genome-wide association and follow-up replication studies identified *ADAMTS18* and *TGFBR3* as bone mass candidate genes in different ethnic groups. *Am. J. Hum. Genet.* **84**, 388–398 (2009).
25. Zheng, H. F. et al. Whole-genome sequencing identifies *EN1* as a determinant of bone density and fracture. *Nature* **526**, 112–117 (2015). **By using whole-genome sequencing, this study identified a low-frequency non-coding variant near the novel locus (*EN1*) associated with lumbar spine BMD and the risk of fracture**.
26. Rivadeneira, F. et al. Twenty bone-mineral-density loci identified by large-scale meta-analysis of genome-wide association studies. *Nat. Genet.* **41**, 1199–1206 (2009).
27. Richards, J. B. et al. Collaborative meta-analysis: associations of 150 candidate genes with osteoporosis and osteoporotic fracture. *Ann. Intern. Med.* **151**, 528–537 (2009).
28. Estrada, K. et al. Genome-wide meta-analysis identifies 56 bone mineral density loci and reveals 14 loci associated with risk of fracture. *Nat. Genet.* **44**, 491–501 (2012).
29. Moayyeri, A. et al. Genetic determinants of heel bone properties: genome-wide association meta-analysis and replication in the GEFOS/GENOMOS consortium. *Hum. Mol. Genet.* **23**, 3054–3068 (2014).
30. Medina-Gomez, C. et al. Life-course genome-wide association study meta-analysis of total body bmd and

assessment of age-specific effects. *Am. J. Hum. Genet.* **102**, 88–102 (2018).

31. Parikka, V. et al. Estrogen responsiveness of bone formation in vitro and altered bone phenotype in aged estrogen receptor-α-deficient male and female mice. *Eur. J. Endocrinol.* **152**, 301–314 (2005).
32. Maatta, J. A. et al. Inactivation of estrogen receptor α in bone-forming cells induces bone loss in female mice. *FASEB J.* **27**, 478–488 (2013).
33. Crockett, J. C., Rogers, M. J., Coxon, F. P., Hocking, L. J. & Helfrich, M. H. Bone remodelling at a glance. *J. Cell Sci.* **124**, 991–998 (2011).
34. Collins, R. What makes UK Biobank special? *Lancet* **379**, 1173–1174 (2012).
35. Kemp, J. P. et al. Identification of 153 new loci associated with heel bone mineral density and functional involvement of *GPC6* in osteoporosis. *Nat. Genet.* **49**, 1468–1475 (2017).
36. Riggs, B. L., Khosla, S. & Melton, L. J. 3rd. A unitary model for involutional osteoporosis: estrogen deficiency causes both type I and type II osteoporosis in postmenopausal women and contributes to bone loss in aging men. *J. Bone Miner. Res.* **13**, 763–773 (1998).
37. Koller, D. L. et al. Meta-analysis of genome-wide studies identifies *WNT16* and *ESR1* SNPs associated with bone mineral density in premenopausal women. *J. Bone Miner. Res.* **28**, 547–558 (2013).
38. Wang, C. et al. Susceptibility genes for osteoporotic fracture in postmenopausal Chinese women. *J. Bone Miner. Res.* **27**, 2582–2591 (2012).
39. Zheng, H. F. et al. *WNT16* influences bone mineral density, cortical bone thickness, bone strength, and osteoporotic fracture risk. *PLOS Genet.* **8**, e1002745 (2012).
40. Simonet, W. S. et al. Osteoprotegerin: a novel secreted protein involved in the regulation of bone density. *Cell* **89**, 309–319 (1997).
41. Xiong, L. et al. Lrp4 in osteoblasts suppresses bone formation and promotes osteoclastogenesis and bone resorption. *Proc. Natl Acad. Sci. USA* **112**, 3487–3492 (2015).
42. Kaprio, J. Twins and the mystery of missing heritability: the contribution of gene-environment interactions. *J. Intern. Med.* **272**, 440–448 (2012).
43. Yang, T. L. et al. Genome-wide copy-number-variation study identified a susceptibility gene, *UGT2B17*, for osteoporosis. *Am. J. Hum. Genet.* **83**, 663–674 (2008).
   The first study to investigate the contribution of copy number variation to osteoporosis identified a deletion variant of *UGT2B17* associated with osteoporotic fracture.
44. Zhu, W. et al. Gene-based GWAS analysis for consecutive studies of GEFOS. *Osteoporos. Int.* **29**, 2645–2658 (2018).
45. Zhang, L. S. et al. Pathway-based genome-wide association analysis identified the importance of regulation-of-autophagy pathway for ultradistal radius BMD. *J. Bone Miner. Res.* **25**, 1572–1580 (2010).
46. Huang, H., Chanda, P., Alonso, A., Bader, J. S. & Arking, D. E. Gene-based tests of association. *PLOS Genet.* **7**, e1002177 (2011).
47. Liu, Y. Z. et al. Powerful bivariate genome-wide association analyses suggest the *Sox6* gene influencing both obesity and osteoporosis phenotypes in males. *PLOS ONE* **4**, e6827 (2009).
48. Medina-Gomez, C. et al. Bivariate genome-wide association meta-analysis of pediatric musculoskeletal traits reveals pleiotropic effects at the *SREBF1*/*TOM1L2* locus. *Nat. Commun.* **8**, 121 (2017).
49. Pei, Y. F. et al. Joint association analysis identified 18 new loci for bone mineral density. *J. Bone Miner. Res.* **34**, 1086–1094 (2019).
50. Greenbaum, J. et al. Increased detection of genetic loci associated with risk predictors of osteoporotic fracture using a pleiotropic cFDR method. *Bone* **99**, 62–68 (2017).
51. Calabrese, G. M. et al. Integrating GWAS and co-expression network data identifies bone mineral density genes *SPTBN1* and *MARK3* and an osteoblast functional module. *Cell Syst.* **4**, 46–59.e4 (2017).
52. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An expanded view of complex traits: from polygenic to omnigenic. *Cell* **169**, 1177–1186 (2017).
53. Lee, S. H. et al. Prediction of future osteoporotic fracture occurrence by genetic profiling: a 6-year follow-up observational study. *J. Clin. Endocrinol. Metab.* **101**, 1215–1224 (2016).
54. Ho-Le, T. P., Center, J. R., Eisman, J. A., Nguyen, H. T. & Nguyen, T. V. Prediction of bone mineral density and fragility fracture by genetic profiling. *J. Bone Miner. Res.* **32**, 285–293 (2017).
55. Kim, S. K. Identification of 613 new loci associated with heel bone mineral density and a polygenic risk score for bone mineral density, osteoporosis and fracture. *PLOS ONE* **13**, e0200785 (2018).
56. Nelson, M. R. et al. The support of human genetic evidence for approved drug indications. *Nat. Genet.* **47**, 856–860 (2015).
57. Xia, W. F. et al. Swedish mutant APP suppresses osteoblast differentiation and causes osteoporotic deficit, which are ameliorated by N-acetyl-L-cysteine. *J. Bone Miner. Res.* **28**, 2122–2135 (2013).
58. Cornelius, C. et al. Osteoporosis and Alzheimer pathology: role of cellular stress response and hormetic redox signaling in aging and bone remodeling. *Front. Pharmacol.* **5**, 120 (2014).
59. Sato, T. et al. Donepezil prevents RANK-induced bone loss via inhibition of osteoclast differentiation by downregulating acetylcholinesterase. *Heliyon* **1**, e00013 (2015).
60. Karasik, D. et al. Heritability and genetic correlations for bone microarchitecture: the framingham study families. *J. Bone Miner. Res.* **32**, 106–114 (2017).
61. Trajanoska, K. et al. Assessment of the genetic and clinical determinants of fracture risk: genome wide association and mendelian randomisation study. *BMJ* **362**, k3225 (2018).
62. Nielson, C. M. et al. Novel genetic variants associated with increased vertebral volumetric BMD, reduced vertebral fracture risk, and increased expression of SLC1A3 and EPHB2. *J. Bone Miner. Res.* **31**, 2085–2097 (2016).
63. Guo, Y. et al. Genome-wide association study identifies *ALDH7A1* as a novel susceptibility gene for osteoporosis. *PLOS Genet.* **6**, e1000806 (2010).
64. Greene, C. S., Penrod, N. M., Williams, S. M. & Moore, J. H. Failure to replicate a genetic association may provide important clues about genetic architecture. *PLOS ONE* **4**, e5639 (2009).
65. Kraft, P., Zeggini, E. & Ioannidis, J. P. Replication in genome-wide association studies. *Stat. Sci.* **24**, 561–573 (2009).
66. Liu, Y. J., Zhang, L., Pei, Y. F., Papasian, C. J. & Deng, H. W. On genome-wide association studies and their meta-analyses: lessons learned from osteoporosis studies. *J. Clin. Endocrinol. Metab.* **98**, E1278–E1282 (2013).
67. Kichaev, G. et al. Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLOS Genet.* **10**, e1004722 (2014).
68. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
69. Benner, C. et al. FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493–1501 (2016).
70. Greenbaum, J. & Deng, H. W. A statistical approach to fine mapping for the identification of potential causal variants related to bone mineral density. *J. Bone Miner. Res.* **32**, 1651–1658 (2017).
71. Benner, C. et al. Prospects of fine-mapping trait-associated genomic regions by using summary statistics from genome-wide association studies. *Am. J. Hum. Genet.* **101**, 539–551 (2017).
72. Liu, Y. Z. et al. A novel pathophysiological mechanism for osteoporosis suggested by an in vivo gene expression study of circulating monocytes. *J. Biol. Chem.* **280**, 29011–29016 (2005).
73. Mohan, S., Hu, Y. & Edderkaoui, B. Chemokine receptor 3 is a negative regulator of trabecular bone mass in female mice. *J. Cell. Biochem.* **120**, 13974–13984 (2019).
74. Fitzpatrick, L. A. et al. Targeted deletion of histidine decarboxylase gene in mice increases bone formation and protects against ovariectomy-induced bone loss. *Proc. Natl Acad. Sci. USA* **100**, 6027–6032 (2003).
75. Takuma, A. et al. Dexamethasone enhances osteoclast formation synergistically with transforming growth factor-β by stimulating the priming of osteoclast progenitors for differentiation into osteoclasts. *J. Biol. Chem.* **278**, 44667–44674 (2003).
76. Eskildsen, T. et al. MicroRNA-138 regulates osteogenic differentiation of human stromal (mesenchymal) stem cells in vivo. *Proc. Natl Acad. Sci. USA* **108**, 6139–6144 (2011).
77. Yang, L. et al. The long non-coding RNA-ORLNC1 regulates bone mass by directing mesenchymal stem cell fate. *Mol. Ther.* **27**, 394–410 (2019).
78. Zhao, K. et al. Hsa_Circ_0001275: a potential novel diagnostic biomarker for postmenopausal osteoporosis. *Cell Physiol. Biochem.* **46**, 2508–2516 (2018).
79. Vidal, C., Cachia, A. & Xuereb-Anastasi, A. Effects of a synonymous variant in exon 9 of the CD44 gene on pre-mRNA splicing in a family with osteoporosis. *Bone* **45**, 736–742 (2009).
80. Scotti, M. M. & Swanson, M. S. RNA mis-splicing in disease. *Nat. Rev. Genet.* **17**, 19–32 (2016).
81. Marini, F., Cianferotti, L. & Brandi, M. L. Epigenetic mechanisms in bone biology and osteoporosis: can they drive therapeutic choices? *Int. J. Mol. Sci.* **17**, E1329 (2016).
82. van Meurs, J. B., Boer, C. G., Lopez-Delgado, L. & Riancho, J. A. Role of epigenomics in bone and cartilage disease. *J. Bone Miner. Res.* **34**, 215–230 (2019).
83. Reppe, S. et al. Methylation of bone *SOST*, its mRNA, and serum sclerostin levels correlate strongly with fracture risk in postmenopausal women. *J. Bone Miner. Res.* **30**, 249–256 (2015).
84. Li, X. et al. Sclerostin binds to LRP5/6 and antagonizes canonical Wnt signaling. *J. Biol. Chem.* **280**, 19883–19887 (2005).
85. Cao, Y. & Wang, B. Expression of sclerostin in osteoporotic fracture patients is associated with DNA methylation in the CpG island of the *SOST* gene. *Int. J. Genomics* **2019**, 7076513 (2019).
86. Delgado-Calle, J. et al. Genome-wide profiling of bone reveals differentially methylated regions in osteoporosis and osteoarthritis. *Arthritis Rheum.* **65**, 197–205 (2013).
   The first epigenome-wide association study of human bone detected 241 differentially methylated CpG sites in femoral head trabecular bone specimens between 27 patients with osteoporotic hip fractures and 26 patients with hip osteoarthritis.
87. Reppe, S. et al. Distinct DNA methylation profiles in bone and blood of osteoporotic and healthy postmenopausal women. *Epigenetics* **12**, 674–687 (2017).
88. Morris, J. A. et al. Epigenome-wide association of DNA methylation in whole blood with bone mineral density. *J. Bone Miner. Res.* **32**, 1644–1650 (2017).
89. Fernandez-Rebollo, E. et al. Primary osteoporosis is not reflected by disease-specific DNA methylation or accelerated epigenetic age in blood. *J. Bone Miner. Res.* **33**, 356–361 (2018).
90. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
91. Roadmap Epigenomics Consortium et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
92. Deng, F. Y. et al. Proteomic analysis of circulating monocytes in Chinese premenopausal females with extremely discordant bone mineral density. *Proteomics* **8**, 4259–4272 (2008).
93. Deng, F. Y. et al. Peripheral blood monocyte-expressed *ANXA2* gene is involved in pathogenesis of osteoporosis in humans. *Mol. Cell Proteom.* **10**, M111.011700 (2011).
94. Deng, F. Y. et al. Is GSN significant for hip BMD in female Caucasians? *Bone* **63**, 69–75 (2014).
95. Zhou, X. et al. *Anxa2* attenuates osteoblast growth and is associated with hip BMD and osteoporotic fracture in Chinese elderly. *PLOS ONE* **13**, e0194781 (2018).
96. Zeng, Y. et al. Mass spectrometry based proteomics profiling of human monocytes. *Protein Cell* **8**, 123–133 (2017).
   The first comprehensive proteome knowledgebase for human monocytes was developed in 2017; it involves a total of 2,237 unique protein-encoding genes and provides a reference map for future in-depth research on monocyte biology and osteoporosis.
97. Bhattacharyya, S., Siegel, E. R., Achenbach, S. J., Khosla, S. & Suva, L. J. Serum biomarker profile associated with high bone turnover and BMD in postmenopausal women. *J. Bone Miner. Res.* **23**, 1106–1117 (2008).
98. Deng, F. Y. et al. An integrative study ascertained *SOD2* as a susceptibility gene for osteoporosis in Chinese. *J. Bone Miner. Res.* **26**, 2695–2701 (2011).
99. Zhang, L. et al. Network-based proteomic analysis for postmenopausal osteoporosis in Caucasian females. *Proteomics* **16**, 12–28 (2016).
100. Zeng, Y. et al. Quantitative proteomics and integrative network analysis identified novel genes and pathways related to osteoporosis. *J. Proteomics* **142**, 45–52 (2016).

101. Arasu, A. et al. Serum sclerostin and risk of hip fracture in older Caucasian women. *J. Clin. Endocrinol. Metab.* **97**, 2027–2032 (2012).
102. Johnson, C. H., Ivanisevic, J. & Siuzdak, G. Metabolomics: beyond biomarkers and towards mechanisms. *Nat. Rev. Mol. Cell Biol.* **17**, 451–459 (2016).
103. Wu, Q. Y. et al. Long non-coding RNAs: a new regulatory code for osteoporosis. *Front. Endocrinol.* **9**, 587 (2018).
104. Ma, B. et al. Metabolomic profiles delineate signature metabolic shifts during estrogen deficiency-induced bone loss in rat by GC-TOF/MS. *PLOS ONE* **8**, e54965 (2013).
105. You, Y. S. et al. Association between the metabolome and low bone mineral density in Taiwanese women determined by ^{1}H NMR spectroscopy. *J. Bone Miner. Res.* **29**, 212–222 (2014).
The first metabolomics study of osteoporosis in humans compared high and low BMD groups and reported four distinguishing metabolites: lactate, acetone, acetate and glutamine.
106. Miyamoto, T. et al. Metabolomics-based profiles predictive of low bone mass in menopausal women. *Bone Rep.* **9**, 11–18 (2018).
107. Zhao, Q. et al. Metabolomic profiles associated with bone mineral density in US Caucasian women. *Nutr. Metab.* **15**, 57 (2018).
108. Yu, L. et al. Association between metabolic profiles in urine and bone mineral density of pre- and postmenopausal Chinese women. *Menopause* **26**, 94–102 (2019).
109. Cabrera, D. et al. Association of plasma lipids and polar metabolites with low bone mineral density in Singaporean-Chinese menopausal women: a pilot study. *Int. J. Environ. Res. Public Health* **15**, E1045 (2018).
110. Moayyeri, A. et al. Metabolomic pathways to osteoporosis in middle-aged women: a genome-metabolome-wide Mendelian randomization study. *J. Bone Miner. Res.* **33**, 643–650 (2018).
111. Liu, L. et al. Assessing the associations of blood metabolites with osteoporosis: a Mendelian randomization study. *J. Clin. Endocrinol. Metab.* **103**, 1850–1855 (2018).
112. Schrimpe-Rutledge, A. C., Codreanu, S. G., Sherrod, S. D. & McLean, J. A. Untargeted metabolomics strategies — challenges and emerging directions. *J. Am. Soc. Mass Spectrom.* **27**, 1897–1905 (2016).
113. Ding, C., Cicuttini, F. & Jones, G. Tibial subchondral bone size and knee cartilage defects: relevance to knee osteoarthritis. *Osteoarthritis Cartilage* **15**, 479–486 (2007).
114. Chou, C. H. et al. Genome-wide expression profiles of subchondral bone in osteoarthritis. *Arthritis Res. Ther.* **15**, R190 (2013).
115. Jaffe, A. E. & Irizarry, R. A. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biol.* **15**, R31 (2014).
116. Teschendorff, A. E. & Zheng, S. C. Cell-type deconvolution in epigenome-wide association studies: a review and recommendations. *Epigenomics* **9**, 757–768 (2017).
117. Maurano, M. T. et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
118. Schaid, D. J., Chen, W. & Larson, N. B. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat. Rev. Genet.* **19**, 491–504 (2018).
119. Boyle, A. P. et al. Annotation of functional variation in personal genomes using regulomeDB. *Genome Res.* **22**, 1790–1797 (2012).
120. Guo, Y. et al. Integrating epigenomic elements and gwass identifies bdnf gene affecting bone mineral density and osteoporotic fracture risk. *Sci. Rep.* **6**, 30558 (2016).
121. Yao, S. et al. Regulatory element-based prediction identifies new susceptibility regulatory variants for osteoporosis. *Hum. Genet.* **136**, 963–974 (2017).
122. Grundberg, E. et al. Population genomics in a disease targeted primary cell model. *Genome Res.* **19**, 1942–1952 (2009).
123. Mullin, B. H. et al. Expression quantitative trait locus study of bone mineral density gwas variants in human osteoclasts. *J. Bone Miner. Res.* **33**, 1044–1051 (2018).
124. Chen, X. F. et al. An osteoporosis risk SNP at 1p36.12 acts as an allele-specific enhancer to modulate LINC00339 expression via long-range loop formation. *Am. J. Hum. Genet.* **102**, 776–793 (2018).
This paper provides a mechanistic explanation for the influence of the 1p36.12 locus on human osteoporosis, using multiple omics technologies to bridge GWAS results to physiology.
125. Meng, X. H. et al. Integration of summary data from GWAS and eQTL studies identified novel causal BMD genes with functional predictions. *Bone* **113**, 41–48 (2018).
126. Giambartolomei, C. et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLOS Genet.* **10**, e1004383 (2014).
127. Hormozdiari, F. et al. Colocalization of GWAS and eQTL signals detects target genes. *Am. J. Hum. Genet.* **99**, 1245–1260 (2016).
128. Chesi, A. et al. Genome-scale Capture C promoter interactions implicate effector genes at GWAS loci for bone mineral density. *Nat. Commun.* **10**, 1260 (2019).
129. Zhu, D. L. et al. Multiple functional variants at 13q14 Risk locus for osteoporosis regulate RANKL expression through long-range super-enhancer. *J. Bone Miner. Res.* **33**, 1335–1346 (2018).
130. Wang, Q. et al. A Bayesian framework that integrates multi-omics data and gene networks predicts risk genes from schizophrenia GWAS data. *Nat. Neurosci.* **22**, 691–699 (2019).
131. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).
132. Lusis, A. J. et al. The Hybrid Mouse Diversity Panel: a resource for systems genetics analyses of metabolic and cardiovascular traits. *J. Lipid Res.* **57**, 925–942 (2016).
133. Calabrese, G. et al. Systems genetic analysis of osteoblast-lineage cells. *PLOS Genet.* **8**, e1003150 (2012).
134. Chen, Y. C. et al. Integrative analysis of genomics and transcriptome data to identify potential functional genes of BMDs in females. *J. Bone Miner. Res.* **31**, 1041–1049 (2016).
135. Chenu, C., Serre, C. M., Raynal, C., Burt-Pichat, B. & Delmas, P. D. Glutamate receptors are expressed by bone cells and are involved in bone resorption. *Bone* **22**, 295–299 (1998).
136. Zhang, J. G. et al. Integrative analysis of transcriptomic and epigenomic data to reveal regulation patterns for BMD variation. *PLOS ONE* **10**, e0138524 (2015).
137. Al-Barghouthi, B. M. & Farber, C. R. Dissecting the genetics of osteoporosis using systems approaches. *Trends Genet.* **35**, 55–67 (2019).
138. Jinek, M. et al. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816–821 (2012).
139. Duan, X. et al. Deficiency of *ATP6V1H* causes bone loss by inhibiting bone resorption and bone formation through the TGF-β1 pathway. *Theranostics* **6**, 2183–2195 (2016).
140. Saito, A. et al. Targeted reversion of induced pluripotent stem cells from patients with human cleidocranial dysplasia improves bone regeneration in a rat calvarial bone defect model. *Stem Cell Res. Ther.* **9**, 12 (2018).
141. Freudenthal, B. et al. Rapid phenotyping of knockout mice to identify genetic determinants of bone strength. *J. Endocrinol.* **231**, R31–R46 (2016).
142. Brommage, R. et al. High-throughput screening of mouse gene knockouts identifies established and novel skeletal phenotypes. *Bone Res.* **2**, 14034 (2014).
143. Chappell, L., Russell, A. J. C. & Voet, T. Single-cell (multi)omics technologies. *Annu. Rev. Genomics Hum. Genet.* **19**, 15–41 (2018).
144. Chan, C. K. F. et al. Identification of the human skeletal stem cell. *Cell* **175**, 43–56.e21 (2018).
145. Santiago-Algarra, D., Dao, L. T. M., Pradel, L., Espana, A. & Spicuglia, S. Recent advances in high-throughput approaches to dissect enhancer function. *F1000Res* **6**, 939 (2017).
146. Chen, Y. C., Greenbaum, J., Shen, H. & Deng, H. W. Association between gut microbiota and bone health: potential mechanisms and prospective. *J. Clin. Endocrinol. Metab.* **102**, 3635–3646 (2017).
147. Jones, R. M., Mulle, J. G. & Pacifici, R. Osteomicrobiology: the influence of gut microbiota on bone in health and disease. *Bone* **115**, 59–67 (2018).
148. Wang, J. et al. Diversity analysis of gut microbiota in osteoporosis and osteopenia patients. *PeerJ* **5**, e3450 (2017).
149. Goodrich, J. K., Davenport, E. R., Clark, A. G. & Ley, R. E. The relationship between the human genome and microbiome comes into view. *Annu. Rev. Genet.* **51**, 413–433 (2017).
150. Timpson, N. J., Greenwood, C. M. T., Soranzo, N., Lawson, D. J. & Richards, J. B. Genetic architecture: the shape of the genetic contribution to human traits and disease. *Nat. Rev. Genet.* **19**, 110–124 (2018).
151. Price, A. L., Zaitlen, N. A., Reich, D. & Patterson, N. New approaches to population stratification in genome-wide association studies. *Nat. Rev. Genet.* **11**, 459–463 (2010).
152. Zhang, L. et al. A new method for estimating effect size distribution and heritability from genome-wide association summary results. *Hum. Genet.* **135**, 171–184 (2016).
153. Wright, N. C. et al. The recent prevalence of osteoporosis and low bone mass in the United States based on bone mineral density at the femoral neck or lumbar spine. *J. Bone Miner. Res.* **29**, 2520–2526 (2014).
154. International Osteoporosis Foundation. Broken bones, broken lives: a roadmap to solve the fragility fracture crisis in Europe (IOF, 2018).
155. Chen, P., Li, Z. & Hu, Y. Prevalence of osteoporosis in China: a meta-analysis and systematic review. *BMC Public Health* **16**, 1039 (2016).
156. Lee, K. S., Bae, S. H., Lee, S. H., Lee, J. & Lee, D. R. New reference data on bone mineral density and the prevalence of osteoporosis in Korean adults aged 50 years or older: the Korea National Health and Nutrition Examination Survey 2008-2010. *J. Korean Med. Sci.* **29**, 1514–1522 (2014).
157. Wade, S. W., Strader, C., Fitzpatrick, L. A., Anthony, M. S. & O'Malley, C. D. Estimating prevalence of osteoporosis: examples from industrialized countries. *Arch. Osteoporos.* **9**, 182 (2014).

Acknowledgements

T.L.Y. and S.S.D. acknowledge the support of the National Natural Science Foundation of China (31771399, 31970569 and 81573241), the Innovative Talent Promotion Plan of Shaanxi Province for Young Sci-Tech New Star (2018KJXX-010) and the special guidance funds for the construction of world-class universities (disciplines) and characteristic development in central universities. H.S. and H.W.D. acknowledge the support of grants from the National Institutes of Health (R01AR059781, P20GM109036, R01MH107354, R01MH104680, R01GM109068, R01AR069055, U19AG055373, R01DK115679), the Edward G. Schlieder Endowment and the Drs. W. C. Tsai and P. T. Kung Professorship in Biostatistics from Tulane University.

Author contributions

H-W.D., T-L.Y., H.S., S-S.D., L.Z., F-Y.D and Q.Z. researched data for the article, made substantial contributions to the discussion of content, and contributed to the writing and review/editing of the manuscript before submission. A.L. researched data for the article and contributed to the writing and review/editing of the manuscript before submission.

Competing interests

The authors declare no competing interests.

Peer review information

Nature Reviews Endocrinology thanks D. Karasik, J. Tobias and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Supplementary information

Supplementary information is available for this paper at https://doi.org/10.1038/s41574-019-0282-7.

RELATED LINKS

FINEMAP: http://www.christianbenner.com/
GCTA-COJO: https://cnsgenomics.com/software/gcta/#COJO
International Mouse Phenotyping Consortium: http://www.mousephenotype.org/
Mouse Genome Informatics: http://www.informatics.jax.org/
PAINTOR: https://github.com/gkichaev/PAINTOR_V3.0
The Origins of Bone and Cartilage Disease project: http://www.boneandcartilage.com/
UK Biobank: https://www.ukbiobank.ac.uk/

© Springer Nature Limited 2019

View publication stats

Comparative genomic analysis of the principal *Cryptosporidium* species that infect humans

Laura M. Arias-Agudelo[1], Gisela Garcia-Montoya[1], Felipe Cabarcas[1,2], Ana L. Galvan-Diaz[3] and Juan F. Alzate[1]

[1] Centro Nacional de Secuenciación Genómica - CNSG, Sede de Investigación Universitaria - SIU, Departamento de Microbiología y Parasitología, Facultad de Medicina, Universidad de Antioquia, Medellin, Antioquia, Colombia

[2] Grupo SISTEMIC, Departamento de Ingeniería Electrónica, Facultad de Ingeniería, Universidad de Antioquia, Medellin, Antioquia, Colombia

[3] Grupo de Microbiología ambiental. Escuela de Microbiología, Universidad de Antioquia, Medellin, Antioquia, Colombia

ABSTRACT

Cryptosporidium parasites are ubiquitous and can infect a broad range of vertebrates and are considered the most frequent protozoa associated with waterborne parasitic outbreaks. The intestine is the target of three of the species most frequently found in humans: *C. hominis*, *C. parvum*, and. *C. meleagridis*. Despite the recent advance in genome sequencing projects for this apicomplexan, a broad genomic comparison including the three species most prevalent in humans have not been published so far. In this work, we downloaded raw NGS data, assembled it under normalized conditions, and compared 23 publicly available genomes of *C. hominis*, *C. parvum*, and *C. meleagridis*. Although few genomes showed highly fragmented assemblies, most of them had less than 500 scaffolds and mean coverage that ranged between 35X and 511X. Synonymous single nucleotide variants were the most common in *C. hominis* and *C. meleagridis*, while in *C. parvum*, they accounted for around 50% of the SNV observed. Furthermore, deleterious nucleotide substitutions common to all three species were more common in genes associated with DNA repair, recombination, and chromosome-associated proteins. Indel events were observed in the 23 studied isolates that spanned up to 500 bases. The highest number of deletions was observed in *C. meleagridis*, followed by *C. hominis*, with more than 60 species-specific deletions found in some isolates of these two species. Although several genes with indel events have been partially annotated, most of them remain to encode uncharacterized proteins.

Subjects Bioinformatics, Genomics, Microbiology, Parasitology, Taxonomy

Keywords *Cryptosporidium*, Molecular biology, Computational biology, High-throughput nucleotide sequencing, Whole-genome sequencing, Genomics, Genetic variation, comparative genomics, NGS, De novo genome assembly

INTRODUCTION

Cryptosporidium is a ubiquitous enteric apicomplexan that infects a broad range of vertebrates, including humans and domestic and wild animals (*Khan, Shaik & Grigg, 2017*). It is described as an important cause of chronic diarrhea in AIDS and other

Submitted 18 May 2020
Accepted 11 November 2020
Published 2 December 2020

Corresponding author
Juan F. Alzate,
jfernando.alzate@udea.edu.co

Academic editor
Erika Braga

Additional Information and Declarations can be found on page 19

DOI **10.7717/peerj.10478**

Ⓒ Copyright
2020 Arias-Agudelo et al.

Distributed under
Creative Commons CC-BY 4.0

OPEN ACCESS

How to cite this article Arias-Agudelo LM, Garcia-Montoya G, Cabarcas F, Galvan-Diaz AL, Alzate JF. 2020. Comparative genomic analysis of the principal *Cryptosporidium* species that infect humans. *PeerJ* **8**:e10478 http://doi.org/10.7717/peerj.10478

immunocompromised patients. It is also a cause of death in children under 24 months old, especially in low-income countries (Sow et al., 2016; Troeger et al., 2017). Furthermore, *Cryptosporidium* is the most frequent protozoa associated with waterborne parasitic outbreaks worldwide (Efstratiou, Ongerth & Karanis, 2017).

Cryptosporidium is classified as a gregarine, within its subclass, the Cryptogregaria (Ryan et al., 2016). Thus far, there are at least 39 species established, and more than 30 genotypes (Firoozi et al., 2019). Among them, approximately twenty-one have been found in humans. However, *Cryptosporidium hominis* and *Cryptosporidium parvum* are responsible for more than 90% of the reported human infection cases worldwide (Grinberg & Widmer, 2016; Feng, Ryan & Xiao, 2018). Other species, occasionally described in humans, including *C. meleagridis*, *C. felis*, *C. canis*, *C. ubiquitum*, *C. cuniculus*, *C. viatorum*, *C. muris*, chipmunk genotype I, *C. andersoni*, *C. suis*, *C. bovis*, horse genotype, *C. xiaoi*, skunk genotype, mink genotype, *C. erinacei*, *C. fayeri*, *C. scrofarum* and *C. tyzzeri* (Feng, Ryan & Xiao, 2018). These species and genotypes differ significantly in human infectivity, host range, geographic distribution, and virulence (Camaa et al., 2007; Cama et al., 2008; Adamu et al., 2014; Feng, Ryan & Xiao, 2018). Some species, such as *C. hominis*, have very narrow host ranges, mostly restricted to humans, nonhuman primates, and horses, whereas others, such as *C. parvum*, have a broad host range, infecting ruminants, horses, rodents, and other animals besides humans (Ryan, Fayer & Xiao, 2014). *Cryptosporidium meleagridis*, which is the third most prevalent species in humans, has been described in mammals and birds (Stensvold et al., 2014).

Cryptosporidium molecular phylogeny and evolutionary relationships have been studied through PCR of single or multiple genetic markers (Feng et al., 2007; Xiao, 2010; Xiao & Feng, 2017; Cunha, Peralta & Peralta, 2019). *Cryptosporidium* species identification protocols usually include the amplification and sequencing of the small subunit (SSU) rRNA gene. Additionally, glycoprotein 60 gene (gp60) has been used to study subtypes and intra-species diversity of the genus, leading to the currently accepted classification into gp60 allelic families subtypes (Xiao, 2010). The SSU rDNA gene in *Cryptosporidium* species does not evolve under a neutral model, and its genetic diversity is restricted to a few polymorphic sites (Sulaiman, Lal & Xiao, 2002). Additionally, some species appear to evolve much quicker than others, according to the SSU rDNA locus.

Regarding the gp60 gene, its high genetic diversity has been attributed to the action of positive selective pressure. Because it can be exchanged by genetic recombination, its typing information does not always agree with other loci, especially with some *C. parvum* subtype families (Feng, Ryan & Xiao, 2018). Other loci used in the *Cryptosporidium* typing include coding genes of actin, 70 kDa heat-shock protein (HSP70), and the *Cryptosporidium* oocyst wall protein (COWP) (Cunha, Peralta & Peralta, 2019). Most of these loci do not evolve neutrally, and the rate of evolution may vary among the different species of the genus (Sulaiman, Lal & Xiao, 2002). According to the above, phylogenetic inference based on analysis of one or a few genetic loci, some under selection pressure, might not reflect the real phylogenetic relationships at the whole genome level (Sulaiman, Lal & Xiao, 2002; Morris et al., 2019).

The growing use of whole-genome shotgun sequencing (WGS) and next-generation sequencing (NGS) in the study of *Cryptosporidium* spp. is allowing better phylogenetic and comparative genomic analysis within the genus (*Widmer & Sullivan, 2012*; *Mazurie et al., 2013a*; *Guo et al., 2015*; *Isaza et al., 2015*; *Ifeonu et al., 2016*; *Beser et al., 2017*; *Sikora et al., 2017a*; *Feng et al., 2017*; *Khan, Shaik & Grigg, 2017*; *Gilchrist et al., 2018*; *Fan, Feng & Xiao, 2019*; *Xu et al., 2019*; *Nader et al., 2019*). The first *Cryptosporidium* genomes were generated in 2004 using capillary sequencing, belonging to the *C. parvum* (Iowa strain) and *C. hominis* (TU502 strain) species (*Xu et al., 2004*; *Abrahamsen, 2004*). Since then, high-quality (NGS-based) genomes from several subtypes of these species and other human-related species are increasingly available (*Widmer & Sullivan, 2012*; *Mazurie et al., 2013a*; *Guo et al., 2015*; *Beser et al., 2017*; *Feng et al., 2017*; *Sikora et al., 2017a*; *Gilchrist et al., 2018*; *Nader et al., 2019*; *Xu et al., 2019*). Comparative analysis shows a remarkable structural and compositional conservation in genome organization among intestinal *Cryptosporidium* species (*Fan, Feng & Xiao, 2019*). The genome size is near 9.0 Mb in length and is arranged into eight chromosomes, with perfect synteny (no evidence of genome rearrangements), extremely compact coding genes, and a low number of gene introns (*Fan, Feng & Xiao, 2019*). It has been postulated that the phenotypic differences between *Cryptosporidium* species could be associated with minor sequence variations (single nucleotide variants-SNVs and short indels) that can affect expressed proteins or gene regulation patterns (*Guo et al., 2015*; *Feng et al., 2017*; *Sikora et al., 2017a*; *Gilchrist et al., 2018*; *Su et al., 2019*; *Xu et al., 2019*). Nonetheless, several studies have demonstrated events of major insertions and deletions between several species of *Cryptosporidium*, usually involving members of multicopy gene families under positive selection, such as those located near telomeres, like the MEDLE proteins, insulinase-like proteases, and mucin-type glycoproteins. These genes have been associated with the host-parasite interaction (*Guo et al., 2015*; *Feng et al., 2017*; *Gilchrist et al., 2018*; *Xu et al., 2019*).

Molecular phylogenetic strategies have been developed to understand the evolutionary relationships between proteins or genes and help to unravel the evolutionary history of the species. Phylogenetic analysis can also give insights into epidemiological, immunological, and evolutionary processes shaping genetic variation in natural populations of the parasite, and may even have the potential to improve future public health measures (*Baele et al., 2017*; *Theys et al., 2019*). Although phylogenomics studies on apicomplexan parasites are scarce, in models such as Piroplasmida and Haemosporida, they have been useful in the elucidation of the taxonomy and phylogenetic relationships within these protozoa, through the incorporation of a broad number of taxons and DNA datasets (*Cornillot et al., 2012*; *Lack, Reichard & Van Den Bussche, 2012*; *Galen et al., 2018*).

Here we used a comprehensive phylogenomic approach to have a better view of the evolutionary relationships of the three most relevant human infecting species of *Cryptosporidium* protozoan parasites (*C. parvum*, *C. hominis*, and *C. meleagridis*), including genomes with different subtype families. Furthermore, a detailed comparative analysis of the largest indel events detected in these three species is presented, which allowed a more comprehensive view of the gene content differences among these three apicomplexan species.

METHODS

NGS data and assembly

Read sequences were downloaded from the Sequence Read Archive - SRA of the NCBI public database. Reads of 23 genomic projects of three *Cryptosporidium* species were included: *C. parvum, C. hominis,* and *C. meleagridis*. Four genomes of *C. meleagridis*, ten genomes of *C. hominis,* and nine genomes of *C. parvum*, including the *C. parvum* anthroponotic isolates UKP14 and UKP15 were analyzed. The reads were extended using FLASH v1.2.11 (*Magoc & Salzberg, 2011*); then the extended and independent read pairs were assembled with the software SPADES v3.11.1 (*Bankevich et al., 2012*), with default settings and testing k-mers of 33, 55, 77, 99, and 111 bases; then the descriptive statistics of the assembly were calculated with in house Perl scripts. To avoid any possible contamination of the sequences with other species, the assembled contigs were aligned using BLASTN with the reference genomes (chromosomes) of *C. hominis* UdeA01, *C. parvum* Iowa II and *C. tyzzeri* UGA55, downloaded from CryptoDB v43 (*Puiu, 2004*). The contigs with a Bit score value ≥ 300 were kept for further analyses.

Sequencing depth analysis

All the reads were mapped against the *C. parvum* Iowa II reference genome downloaded from CryptoDB v43 and against the *de novo* assembled contigs using BWA (*Li & Durbin, 2010*) with default options. Then, the Samtools-depth tool (*Li et al., 2009*) was used to compute the read depth at all positions, with a maximum coverage depth to 1,000,000. Finally, the mean coverage was estimated in the Linux terminal with an awk formula.

Single nucleotide variants detection

Detection of SNVs was performed aligning with MUMmer v3 the assembled scaffolds for each isolate with the *C. parvum* Iowa II reference genome. Then, the MUMmer function show-SNPs, and dnaDIFF tools were used (*Delcher et al., 2002*).

Phylogenomic analysis

The phylogenomic analysis was performed based on the SNVs found within the 24 selected *Cryptosporidium* genomes. A matrix comprising all the SNVs detected was constructed and loaded into the program IQ-TREE v1.6.12 (*Trifinopoulos et al., 2016*). A maximum-likelihood (ML) tree (*Felsenstein, 1981*) was built using this phylogenomic inference software. Modelfinder was used to select the best model of evolution under Bayesian Information Criterion (BIC) (*Schwarz, 1978*). Transversion Model TVM (TVM+F+ASC+R2) was selected as an evolutionary model. Branch support was estimated using 1000 iterations of the Shimodaira-Hasegawa test (SH -aLRT) (*Shimodaira & Hasegawa, 1999*),Shimodaira Hasegawa 1000 pseudoreplicates of ultra-fast bootstrap (*Hoang et al., 2018*). The presented tree is unrooted, and the longest branch was selected as an arbitrary outgroup. The tree was edited using FigTree v1.4.4 (*Andrew, 2018*).

Identification of SNVs in coding regions

For the detection of single nucleotide variants in the CDSs, the reads were mapped to the genome of *C. parvum* Iowa II with the BWA (Burrows-Wheeler Aligner) aligner version

0.7.17 (*Li & Durbin, 2010*). Later, the bam and VCF files were generated with SAMtools and BCFtools, respectively (*Danecek et al., 2011*). Only variants with a Phred quality score $\geq$ 50 (*Holder et al., 2013*) were included. Finally, the SAMtools Depth tool was used to calculate the reading depth at all positions, with a maximum coverage depth of 1,000,000. Subsequently, the mean coverage was estimated with an awk formula. The single nucleotide variants in CDSs were annotated with the effect predictor SIFT4G (sorting intolerant from tolerant) version 3.0 (*Vaser et al., 2016*).

Identification of insertions and deletions

Insertions and deletions –indels - were identified with MUMmer aligner (Maximal Unique Matches) version 3.0 (*Delcher et al., 2002*), as previously described by *Isaza et al. (2015)*, excluding those with a length <50 nucleotides with the software online Assemblytics (*Nattestad & Schatz, 2016*). To perform the functional analyzes, the genome of *C. parvum* Iowa II version 43 deposited in CryptoDB (*Puiu, 2004*) was used as a reference. Variants were initially detected in coding regions, and then genes with shared and species-exclusive deleterious mutations were identified and annotated.

Identification and annotation of genes with deleterious non-synonymous changes

From the prediction obtained with SIFT4G, genes with deleterious mutations in the eight chromosomes were filtered for each of the genomes. These genes were extracted with the SeqSelect.py script, and enrichment analysis in Gene Ontology terms was performed with the EggNOG mapper program version 1.0 (*Huerta-Cepas et al., 2017*); and visualization with WEGO (Web Gene Ontology Annotation Plotting) version 2.0 was done (*Ye et al., 2018*). Then, functional orthology analysis was performed with KEGG on the KAAS server (*Ogata et al., 1999*), selecting the best hits using a bidirectional strategy - BBH (bi-directional best hit). To detect genes with transmembrane domains, the TMHMM version 2.0 server (*Krogh et al., 2001*) was used. Prediction of the genes that code for proteins with classical and non-classical secretion was performed with the online servers SIGNALP version 5.0 (*Almagro Armenteros et al., 2019*) and SECRETOME-P version 2.0 (*Bendtsen et al., 2004*), respectively.

RESULTS

Selected *Cryptosporidium* genomes

To carry out a comprehensive genomic comparative analysis, we chose all the NGS projects of the three main species of *Cryptosporidium* that infect humans (*C. hominis, C. parvum,* and *C. meleagridis*), that have raw sequence data available in public databases. A total of 23 genomes were included: ten isolates of *Cryptosporidium hominis* (UKH1, UKH3, UKH4, UKH5, 37999, TU502_2012, 30976, UdeA01, SWEH2, and SWEH5); nine isolates of *C. parvum* (UKP2, UKP3, UKP4, UKP5, UKP6, UKP7, UKP8, and the *C. parvum* anthroponotic isolates UKP14 and UKP15); and four isolates of *C. meleagridis* (UKMEL1, UKMEL3, UKMEL4, and TU1867). The low representation of *C. meleagridis* genomes is due to the limited number of sequencing projects for this species, so all available genomes in public databases until September 2019 were included. The genomes belong to different gp60 subtypes, and the majority comes from the UK. All the studied genomes come from parasite oocysts isolated from human feces of patients with natural infections, and four of these were maintained through passages in piglets, mice, or chickens. Most genomes were sequenced on Illumina's HiSeq and MiSeq platforms, and only two (*C. hominis* SWEH2 and SWEH5) used the Ion Torrent platform (Table S1).

De novo genome assembly analysis

The raw read data of the 23 selected genomic projects were downloaded from the Sequence Read Archive - SRA of the NCBI public database and assembled with SPAdes. If the isolate had assembly data available, the metrics of the two assemblies were compared, and the contig set that showed the better N50 value was kept for further analysis. Only six genomes of the studied had better N50 values than our assemblies. All of them were deposited at the CryptoDB database v43: *C. hominis*: UKH1, TU502_2012, 30976, 37999, UdeA01, and *C. meleagridis* UKMEL1 (Table S2). Only for three genomes, it was not possible to find a previous assembly data available: *C. hominis* (SWEH2 - SWEH5) and *C. meleagridis* (TU1867). To exclude possible contaminating contigs, BLASTN comparisons with a *Cryptosporidium* genome database (reference genomes of *C. hominis* UdeA01, *C. parvum* Iowa II and *C. tyzzeri* UGA55) were performed, and only those with Bit score value ≥ 300 were kept for further analyses. Selected genomes assemblages' statistics are also shown in Table S2.

Cryptosporidium genome size is close to 9.0 Mb in most studied isolates; only three have genome assemblies below 9.0 Mb; 8.2 Mb, in *C. parvum* anthroponotic UKP14, and 8.8 Mb in both *C. hominis* SWEH2 and SWEH5 isolates. Four genomes had a mean read coverage below 50X, three of which belong to *C. hominis* species (UKH3, SWEH2, and SWEH5) and the *C. parvum* isolate UKP5. The most fragmented genomes were in the *C. parvum* species, isolate UKP3 and the anthroponotic UKP14 with 2,971 and 2,787 contigs, respectively, with coverage above 229X and 69X. All genomes have an average GC content of 30% and ambiguities that do not exceed 0.26% (Table S2).

Single nucleotide variants detection

To identify single nucleotide variants -SNVs- throughout the genomes, inter and intra-species comparisons were made aligning the *de novo* assembled contigs of each isolate against the reference genome of *C. parvum* Iowa II. *Cryptosporidium parvum* isolates aligned more than 99.2% of the contig bases to the Iowa reference, except for the anthroponotic genomes, UKP14 and UKP15, that aligned 97%. *C. hominis* genome data behave similarly, with an aligned ratio that ranges between 99 to 99.38%, except for UKH4 that reached 98%. In *C. meleagridis*, a lower alignment rate was achieved with an overall rate close to 97% (Table S3).

The global nucleotide identity showed the expected results based on the phylogenetic relationship among the three species. While *C. parvum* genome identity ranged between 99.51% and 99.93% within the species, *C. hominis* genomes showed an identity of around 96.8% compared to the Iowa reference genome. *Cryptosporidium meleagridis* confirmed its higher distance with *C. parvum* with a global identity of around 91.5%.

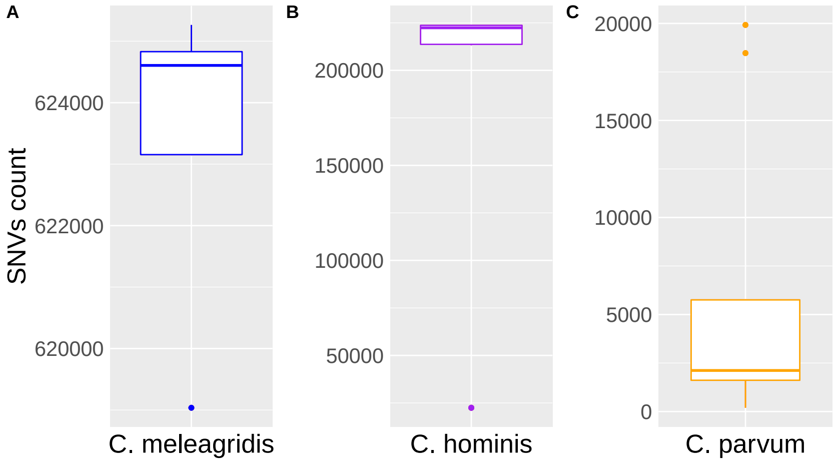


Figure 1 **Accumulation of SNVs in Cryptosporidium species.** Box plots of the *Cryptosporidium* species analyzed and the number of total SNVs detected. The medians of the SNVs found in each species are (A) *C. meleagridis* (624,607), (B) *C. hominis* (223,640), and (C) *C. parvum* (2,116). Outliers in *C. meleagridis* and C. hominis correspond to the UKMEL3, SWEH2, and SWEH5, respectively. Concerning *C. parvum*, outliers correspond to anthroponotic UKP14 and UKP15 isolates.

Full-size 🖼 DOI: 10.7717/peerj.10478/fig-1

Single nucleotide variants within the *C. parvum* genomes ranged between 1,595 and 5,752, except for the anthroponotic isolates that showed more than 18,000 SNVs. When *C. hominis* genomes were compared with the *C. parvum* Iowa II reference, around 220,000 SNVs were detected in each isolate, while in *C. meleagridis* genomes, the number when up to more than 600,000 (Fig. 1).

Phylogenomic analysis

To verify the topology and taxonomic location described in the genus *Cryptosporidium*, the SNV data was used to generate a nucleotide matrix. Then, an unrooted maximum likelihood (ML) tree was constructed, including the 24 evaluated genomes (23 de novo assembled genomes and the reference genome Iowa II). The transversion model (TVM) was selected as the best model of evolution Table S4). The complete matrix comprised 800,861 sites and the best tree had had a likelihood value of −3401691.458. The tree obtained shows three monophyletic clades, following the actual classification scheme, that group the 24 genomes of the three *Cryptosporidium* species included in this work. The branch supports are optimal and have a 100% agreement between SH-aLRT and ultra-fast bootstrap for the main branches, with discrepancies in three internal *C. parvum* nodes and one internal *C. hominis* node (Fig. 2).

Within the *C. parvum* species, the presence of two separate branches with statistical supports was observed, which allowed the segregation of the anthroponotic isolates UKP14 and UKP15 in a separate branch of the zoonotic isolates with a 100% support. Furthermore, it is noteworthy to point out that all the *C. parvum* isolates that belong to the gp60 gene IIa subtype family grouped with a 100% bootstrap, including the isolate Iowa II. In the case of *C. hominis* clade, the isolates seem to segregate with statistical support according to its subtype family, except for isolate 30976. The *C. meleagridis* clade is also supported with

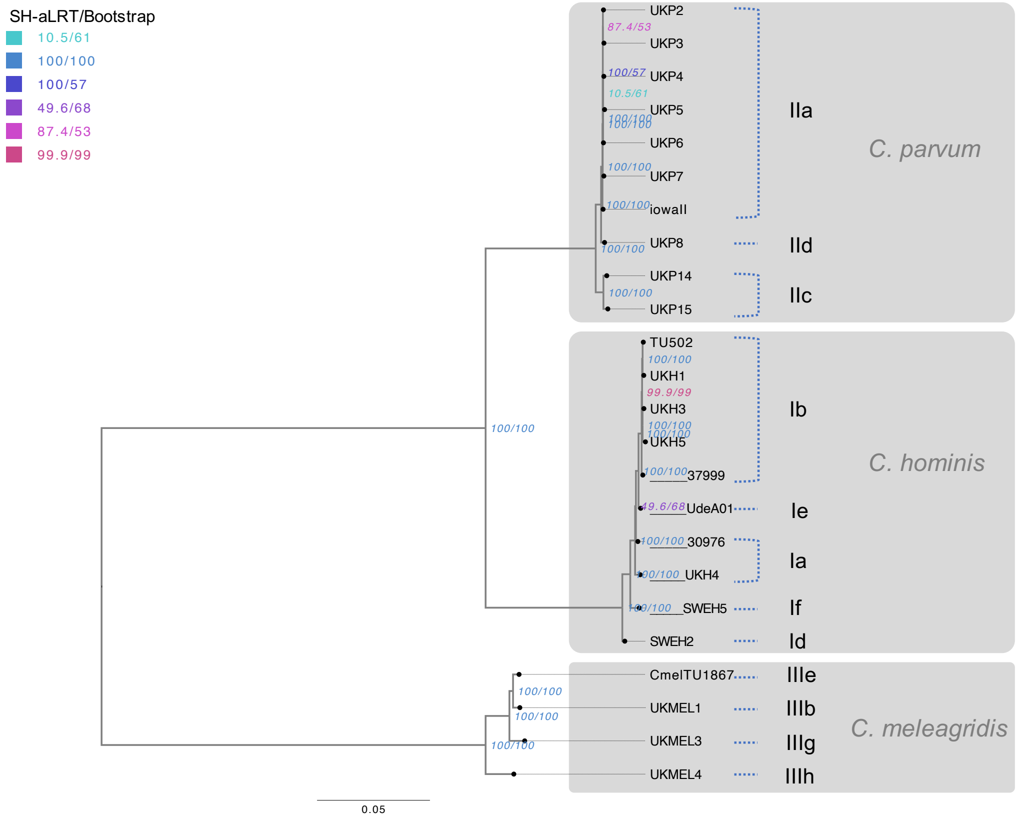

SH-aLRT/Bootstrap
10.5/61
100/100
100/57
49.6/68
87.4/53
99.9/99
UKP2
87.4/53
UKP3
100/57 UKP4
10.5/61
UKP5
100/100
100/100
UKP6
100/100
UKP7
100/100 iowaII
100/100 UKP8
UKP14
100/100
UKP15
IIa
IId
IIc
C. parvum
TU502
100/100
UKH1
99.9/99
UKH3
100/100
100/100
UKH5
100/100 37999
49.6/68 UdeA01
100/100 30976
100/100 UKH4
100/100 SWEH5
SWEH2
Ib
Ie
Ia
If
Id
C. hominis
100/100
CmelTU1867
100/100
UKMEL1
100/100
UKMEL3
100/100
UKMEL4
IIIe
IIIb
IIIg
IIIh
C. meleagridis
0.05

Figure 2 **Phylogenomic analysis of *Cryptosporidium* species.** Maximum Likelihood (ML) tree was based on the SNVs present within the 24 *Cryptosporidium* genomes studied. The presented tree is unrooted, and the longest branch was selected as an arbitrary outgroup. The horizontal scale line represents the number of base substitutions per site analyzed. TVM was the evolutionary model. The supports of the branches are based on an analysis of 1000 **aLRT** replications (%) / ultra-fast **bootstrap** replications (%). Red branches have an aLRT and bootstrap value of 100%. Branches are labeled with the isolate ID and allelic family based on the *gp60* gene.

Full-size DOI: 10.7717/peerj.10478/fig-2

100% bootstrap, and the phylogenetic signal is enough to separate the 4 different isolates according to its subtype family. In general, all genomes were grouped according to the species classification and its gp60 gene subtype family (Fig. 2).

Single nucleotide variants in coding regions

SNVs located in coding regions were identified in all the genomes through the read mapping analysis against the *C. parvum* Iowa II version 43 of CryptoDB. As expected, compared with the *C. parvum* IOWA reference, the genomes of *C. hominis* and *C. meleagridis* showed the highest number of these variants in coding regions with more than 150,000 and 400,000, respectively, and nearly 60% corresponding to synonymous changes (sSNVs). There were no intraspecies differences in the accumulation of variants in coding regions or synonymous and non-synonymous changes in both species. On the contrary, in *C. parvum* isolates, the number of SNVs in CDSs was less than 20,000 in all genomes, with more than 50% corresponding to non-synonymous mutations. The results in *C. parvum* suggest an

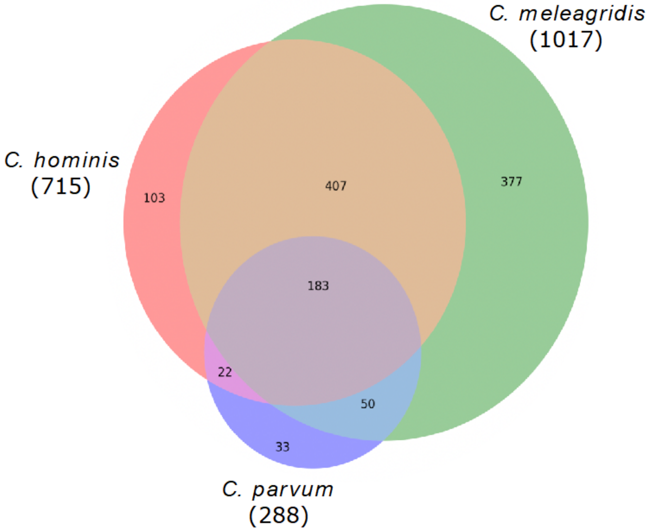


Figure 3 **Genes detected with deleterious mutations in the three *Cryptosporidium* species.** Venn diagram showing the genes with shared and exclusive deleterious mutations in the three *Cryptosporidium* species included in the study.

Full-size 🖼 DOI: 10.7717/peerj.10478/fig-3

intraspecies heterogeneity with values ranging between 840 and 3,476 SNVs in CDSs in zoonotic isolates and around 14,000 in anthroponotic isolates. However, these differences were not statistically significant. After characterizing the synonymous and non-synonymous mutations in coding regions, the genes with nsSNVs were analyzed separately, finding 2,532 in at least one genome of each species, with 110 exclusives for *C. meleagridis*, 11 for *C. hominis,* and 2 for *C. parvum*.

Genes with non-synonymous variants (nsSNVs) were annotated by the SIFT4G effect predictor, and those with non-tolerated changes and possibly associated with deleterious mutations were identified. With this strategy, a total of 1,017 genes were predicted with deleterious mutations in *C. meleagridis*, 715 in *C. hominis,* and 288 in *C. parvum*. Of these, 183 are shared by all three species, being present in at least one genome of each species; 377 are exclusive to *C. meleagridis*; 103 to *C. hominis,* and 33 to *C. parvum* (Fig. 3).

Only 29.5% of the genes annotated with deleterious mutations by SIFT, and shared by the three species, could be annotated with EggNOG mapper. Enrichment analysis in GO terms indicated that they are mainly involved in biological processes and metabolic processes with molecular functions such as catalytic or binding activity.

The functional orthology analysis performed on the KAAS server against the KEGG database had a better performance compared to the previous GO assignment, annotating 46.5% of the shared genes and indicating that most of them are involved in enzymatic processes, DNA repair, recombination, and proteins associated with the chromosomes (Fig. 4). Complementary annotation determined that only 1.1% of the genes encoded proteins secreted by the classical pathway, whereas 10.9% encoded for proteins secreted

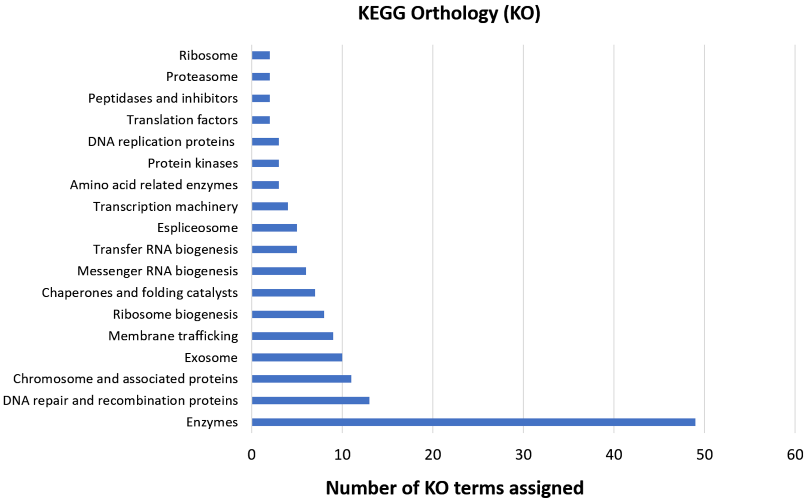


Figure 4 **KEGG orthology analysis of genes with deleterious mutations shared by the three *Cryptosporidium* species.** Assignment of KO terms to genes with deleterious mutations detected in the three *Cryptosporidium* species through the KAAS server in the KEGG database.

Full-size DOI: 10.7717/peerj.10478/fig-4

by non-classical systems. Additionally, 2.3% of the putative encoded proteins exhibited domains with transmembrane helixes.

Regarding genes that carry deleterious mutations within each species (species-specific mutated genes), 29.1 and 28.9% were classified in at least one GO category for *C. hominis* and *C. meleagridis*, respectively. In *C. parvum*, GO terms were assigned only in 3 genes (9.09%), classified exclusively in the cellular component category. The functional orthology analysis again achieved a better result compared to the EggNOG-mapper software, assigning 42.4% in *C. parvum* (Fig. 5A), 60.2% KO in *C. hominis* (Fig. 5B), and 59.1% in *C. meleagridis* (Fig. 5C). The assignment of functional orthologs in metabolic pathways established that in the three species, most of these genes are involved in enzymatic processes. The second most frequently assigned pathway in genes with species-exclusive deleterious mutations was ribosome transfer RNA biogenesis in *C. parvum*, spliceosome complex in *C. hominis*, and biogenesis in *C. meleagridis*.

Indel events analysis

Inter and intraspecies indel events were identified in the genomes through their comparison to the *C. parvum* Iowa II reference genome. For simplicity reasons, insertions and deletions were referred, assuming as reference the *C. parvum* Iowa II genome. Observed insertions were in the range of 50 to 500 nucleotides. In *C. parvum*, the highest number of insertions was detected in the zoonotic isolates UKP3 and UKP4. No statistically significant differences were found in the accumulation of insertions among the studied species. Deletions identified in the genomes had a similar size to that described for the insertions, with the majority falling into the range of 50 to 500 nucleotides. The highest number was observed in the genomes of *C. meleagridis*, followed by *C. hominis*. However, no significant differences were found in the accumulation of deletions among species.

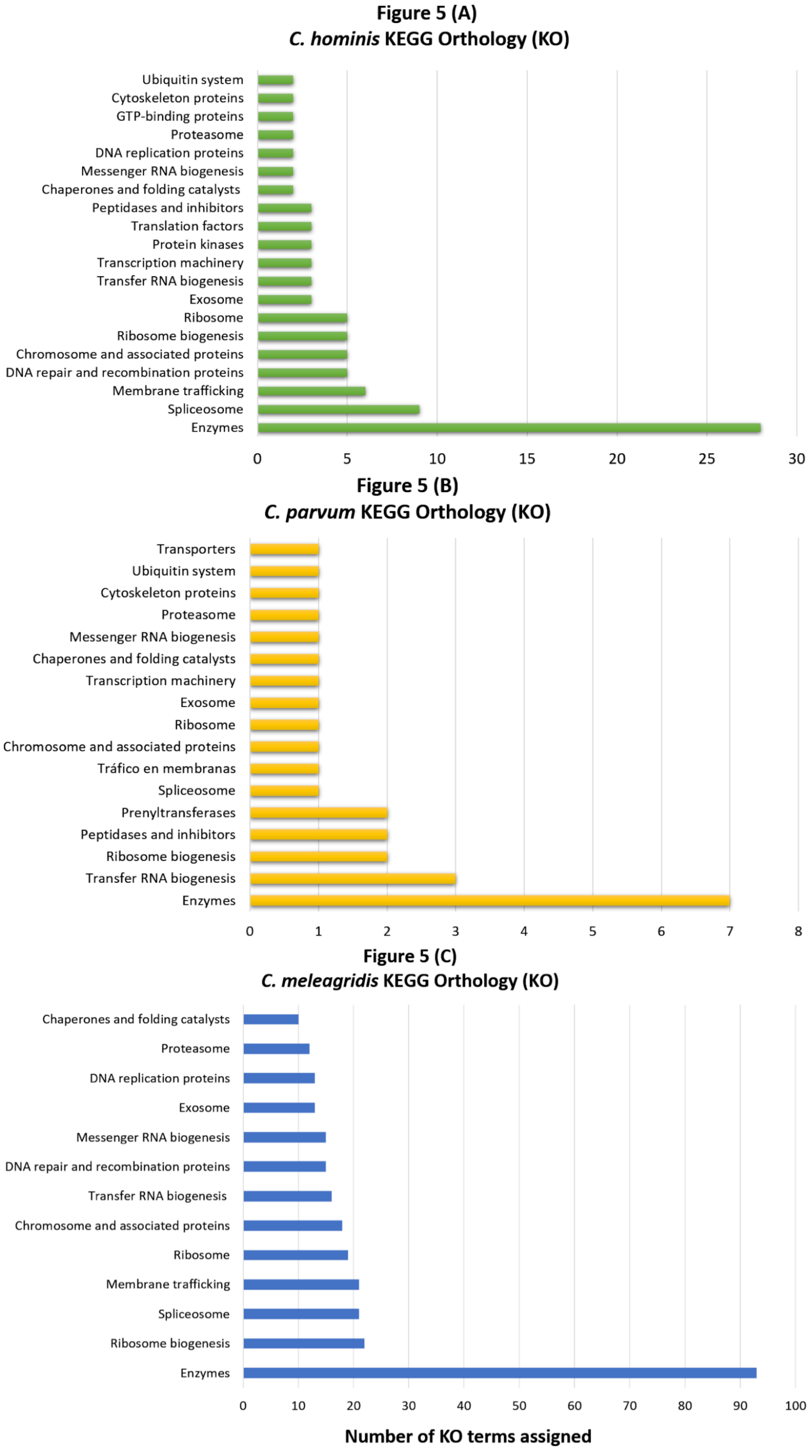


Figure 5 **KEGG orthology analysis of genes with species-exclusive deleterious mutations.** Assignment of KO terms to genes with species-exclusive deleterious mutations via the KAAS server in the KEGG database in *C. hominis* (A), *C. parvum* (B), and *C. meleagridis* (C).

Full-size DOI: 10.7717/peerj.10478/fig-5

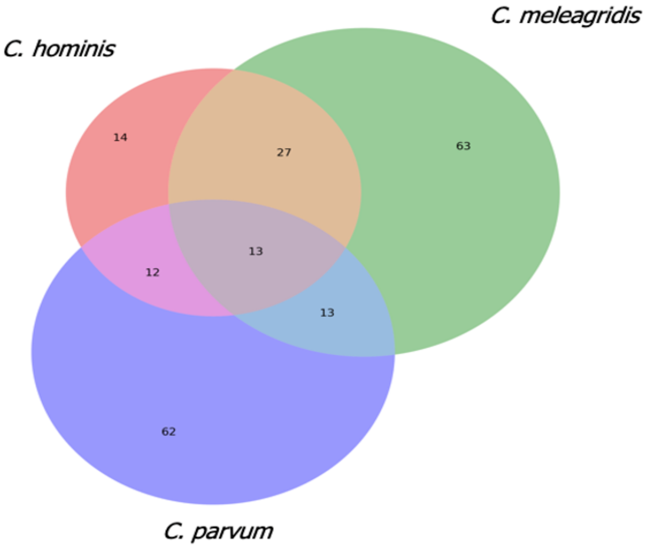


Figure 6 **Analysis of genes with indel events in the three *Cryptosporidium* species.** Venn diagram that shows the genes with indel events that are shared and exclusive in the three *Cryptosporidium* species.
Full-size DOI: 10.7717/peerj.10478/fig-6

Coding genes that presented indels in the three species were identified and annotated, whenever possible. We found 322, 215, and 176 genes with indels in *C. hominis*, *C. parvum*, and *C. meleagridis*, respectively, but only 13 were common to all 23 *Cryptosporidium* isolates. In both *C. meleagridis* and *C. hominis*, deletions were the most frequent indel event, corresponding to 76.1% and 59% of all structural variants, respectively. The *C. hominis* isolates UKH4 was the exception, in which insertions predominated. In contrast, in *C. parvum*, deletions were less common, representing 38% of the structural variants detected in the CDSs.

To gain insights into the gene losses occurring in the 23 isolates, those genes with indel events present in at least two genomes/isolates in each species were characterized. A total of 116 genes were selected in *C. meleagridis*, 100 in *C. parvum*, and 66 in *C. hominis*. Sixty-three were exclusive to *C. meleagridis*, 62 to *C. parvum*, and 14 to *C. hominis* (Fig. 6).

According to the annotation deposited in CryptoDB Version 46, of the 13 genes with insertion and deletion events shared by the three species, 40% code for uncharacterized proteins and the remaining for hydrolases (cgd8_1220), proteins with recognition motifs of RNA (cgd3_4150), mucins (cgd7_4020), among others. Twenty-seven genes with indels were identified in all *C. meleagridis* genomes, with 56% of them corresponding to uncharacterized proteins. Regarding *C. parvum*, insertions in the genes cgd3_190 (involved in the formation of fibrillin) and cgd3_3900 (uncharacterized protein) were present in 56% of the evaluated genomes. An interesting finding was that all *C. hominis* genomes exhibited indels in three genes (orthologs of cgd6_4290, cgd7_420, and cgd7_500); with cgd6_4290 and cgd7_420 being also affected in all genomes of *C. meleagridis* (Table S5).

DISCUSSION

Advances in DNA sequencing technologies and bioinformatics have promoted the routine use of complete genome sequences, revolutionizing the study of both model and non-model organisms, particularly in microbiology (*Young & Gillung, 2020*). Phylogenomic is one of the numerous disciplines that have taken advantage of the progress in NGS technologies, using the massive datasets to infer both phylogenetic relationships between taxa, improving the understanding of molecular evolution, and putative functions for DNA or protein sequences (*Young & Gillung, 2020*).

A phylogenomic analysis is being applied now to validate previous findings with classic single-marker molecular methods and to resolve with unprecedented resolution, the phylogenetic relationships between species and isolates of the human infecting *Cryptosporidium* parasites (*Glaberman et al., 2001*; *Abe & Makino, 2010*; *Abal-Fabeiro et al., 2013*; *Wagnerová et al., 2015*; *Pérez-Cordón et al., 2016*). However, this phylogenomic approach has focused on *C. hominis* and *C. parvum* by studying multiple concatenated loci (*Feng et al., 2017*; *Nader et al., 2019*), or different variable positions that represent below 1% of the genome (*Gilchrist et al., 2018*). To gain insights into the evolutionary relationships in *Cryptosporidium,* a comparative and phylogenomic study was conducted with 23 genomes of the most frequent species in humans: *C. hominis*, *C. parvum*, and *C. meleagridis*.

The sequence identity between *C. hominis* genomes (96.85%) and the reference genome of *C. parvum* Iowa II found in our study confirms the high similarity between both species and agrees with previous findings that report differences of a maximum of 3% (*Xu et al., 2004*; *Mazurie et al., 2013b*; *Zahedi et al., 2013*). These studies describe that both genomes exhibit remarkable structural conservation, and some authors suggest that the phenotypic differences may be due to subtle variations in the sequences of genes that code for the interface proteins between the parasite and its host (*Xu et al., 2004*; *Mazurie et al., 2013b*; *Zahedi et al., 2013*). Regarding *C. parvum*, there was a lower percentage of aligned blocks against the reference genome in the anthroponotic isolates compared to the zoonotic ones, which agrees with other studies that have also described differences in the genomic structure and nucleotide content between zoonotic and anthroponotic isolates of *C. parvum* (*Widmer et al., 2012*; *Fan, Feng & Xiao, 2019*; *Nader et al., 2019*). As it was expected, *C. meleagridis* genomes had a lower global identity against *C. parvum* reference genome than that detected with *C. hominis*, which agrees with the phylogenetic relationships described for the three species using different genetic locus (*Šlapeta, 2013*; *Khan, Shaik & Grigg, 2017*). This finding may be related to the differences in host specificity reported in the three species, where *C. hominis* and *C. parvum* have host ranges limited to mammals, while *C. meleagridis* is described in mammals and birds (*Šlapeta, 2013*; *Khan, Shaik & Grigg, 2017*). It has been suggested that mammals were possibly the original hosts for *C. meleagridis* and that later this species adapted to birds (*Caccio & Widmer, 2014*).

Single nucleotide variants (SNVs) results obtained in this study showed similar behavior in their number and frequency in the genomes of *C. hominis* with around 220,000 variants, compared to the *C. parvum* IOWA reference, results that differ from those reported in previous comparative studies (*Isaza et al., 2015*; *Gilchrist et al., 2018*). Isaza et al. (*Isaza et al., 2015*) identified an average of 43,258 SNVs between the reference genome *C. parvum* Iowa II and the *C. hominis* UdeA01, UKH1, and TU502_2012 isolates (deposited in CryptoDB version 8). One possible explanation of the lower number of SNVs detected in this work could be the different methodological approaches used to identify these variants. Authors only considered those SNVs located at 30 nucleotides or more from regions where the alignment failed, underestimating the total number of SNVs. Additionally, the genomes used in Isaza's study were obtained from the CryptoDB version 8 database, whose genomes had a lower degree of purification than those used in our work (version 43). *Gilchrist et al. (2018)* also found a lower number of SNVs (36,780) in a comparative study of 32 genomes of *C. hominis*. The authors mapped the short reads of the *C. hominis* genomes against the genome of *C. parvum* Iowa, which was sequenced with Pacific Biosciences (PacBio). It could be possible that the assembly of a genome through long sequencing reads can affect the number of variants detected. Although it is not clear how long reads sequencing methods might differ from comparative genomic approaches using short-read data, i.e., Illumina (*DeMaio et al., 2019*), it has been shown that error rates on this platform are much higher than those recorded for Illumina (15% versus 0.1%) and usually is performed at a lower read coverage compare to Illumina projects.

Concerning the single nucleotide variants among the *C. parvum* genomes evaluated in this study, a more significant number of SNVs was identified in the anthroponotic isolates with more than 18,000 compared to the zoonotic isolates, which ranged between 1,595 and 5,752 SNVs. These results are similar to the findings reported by *Widmer et al. (2012)*, who identified about 16,606 SNVs by comparing the genomes of two *C. parvum* isolates with different host ranges, the anthroponotic *C. parvum* TU114 and the reference genome of the zoonotic Iowa II isolate. The differential accumulation of intragenotypic SNVs in *C. parvum* reflects the genetic diversity of this species, which is possibly related to the evolution of the parasite influenced by the selective pressures in both humans and animal hosts. For this reason, it has been postulated that the accumulation of genomic variants could influence the host range (*Weir et al., 2011*; *Blake et al., 2015*; *Grinberg & Widmer, 2016*). Concerning *C. meleagridis*, more than 600,000 SNVs were identified in the genomes compared against the reference genome of *C. parvum*. It is essential to highlight that there are currently no reports on the structural variations like SNVs in this species, being this the first one.

In this work, a comprehensive phylogenomic analysis using more than 800.000 single nucleotide variations detected in 24 genomes of the species *C. parvum*, *C. hominis,* and *C. meleagridis* was done. To our knowledge, this is the most extensive phylogenomic analysis carried out within the genus and one of the largest within the Phylum Apicomplexa. One of the main findings in this study is that, although we didn't select neutral evolving positions within the genomes, there was a strong phylogenetic signal, supported by two statistical tests, that allowed the well-supported segregation for most of the isolates. However, discrepancies were found in three internal *C. parvum* nodes and one in *C. hominis*. It is essential to highlight that the mentioned differences did not alter the global topology of the phylogenomic tree obtained. In related literature, it has been described that the mismatches in the supports of the branches obtained by bootstrapping and probability ratio tests can arise as a consequence of performing the analysis with small samples and with highly heterogeneous nucleotide substitution models (Guindon et al., 2010).

The obtained tree confirms the previously reported topology for these three intestinal species of *Cryptosporidium*, inferred from single-locus phylogenetic studies, and by the use of different loci such as 18S rRNA, gp60, and other polymorphic genes (Ren et al., 2012; Šlapeta, 2013; Nader et al., 2019). *Cryptosporidium meleagridis* is confirmed as the most divergent group among the three studied species (Šlapeta, 2013; Khan, Shaik & Grigg, 2017). At the species rank, the bootstrap and SH-aLRT statistical support was 100%. Furthermore, most of the isolates were grouped according to its *gp60* gene family type, regardless of its geographical origin.

The phylogenomic analysis of the *C. parvum* isolates evidenced the separation of its central clade into two branches with significant statistical support (100%), with zoonotic isolates in one branch and anthroponotic isolates in the other. This finding agrees with that reported by several authors (Widmer et al., 2012; Danišová et al., 2017; Feng et al., 2017; Nader et al., 2019), in which through unilocus phylogenetic analyzes and multiple concatenated loci between anthroponotic and zoonotic isolates, determined that the anthroponotic isolates of *C. parvum* formed a separate group from the zoonotic isolates. Nader et al. (2019), through phylogenetic analysis of neutrally evolving coding loci across 21 *Cryptosporidium* isolates, identified two *C. parvum* lineages with distinct host-specificity, which were designated as *Cryptosporidium parvum parvum (* zoonotic) and *C. p. anthroponosum* (anthroponotic). Additionally, they found that human infective *C. hominis* and *C. parvum* isolates form a distinct superclade along with *C. cuniculus*, another species associated with human infections. Subsequent analysis of high-quality SNPs detected in 16 genomes of the two *Cryptosporidium parvum* subclades, confirming the zoonotic *C. p. parvum* and anthroponotic *C. p. anthroponosum* subspecies designation (Nader et al., 2019).

Feng et al. (2017) evaluated the SNVs accumulation in several genomes of *C. parvum* IIa and IId families, which preferentially infect calves and lambs, respectively, in some European countries. This study revealed that most of the SNVs occur in subtelomeric regions of the chromosomes, with a high percentage located in coding regions of the genome, and near the half being non-synonymous. Additionally, the subtypes evaluated shared more than 50% of SNVs, and phylogenetic analysis of the SNVs data showed a robust separation of IIa sequences and IId sequences, and a high divergence with reference Iowa genome (Feng et al., 2017). These findings agree with the results obtained in this study because SNVs show a concordant relationship with gp60 subfamily typing.

Conversely, Gilchrist et al. (2018) studied the genetic diversity of thirty-two genomes of different *C. hominis* subtypes isolated from children with poor living conditions from Bangladesh. They found 36,780 SNVs that varied between the *C. hominis* isolates, with a homogeneous distribution throughout the genome and only 4% occurring with a frequency greater than 20%. These authors also built a phylogenetic tree based on the SNVs (1,582) found in those genomes, in which no groupings regarding the subtype family was observed, concluding that the use of a single marker (gp60) does not reflect the evolutionary changes of the entire genome and, in turn, confirming the weakness of the typing of unique markers for taxonomic assignments within this genus. Our findings reinforce this argument since an analysis with a more significant number of positions of different isolates considerably improves the resolution power compared to that obtained from unilocus analysis or multiple concatenated loci or partial fragments of the genome. Another aspect that could influence the topology of the phylogenomic tree obtained by Gilchrist et al. is the high rate of recombination on chromosome 6 reported among circulating isolates from endemic countries for *C. hominis* (*Li et al., 2013*; *Zahedi et al., 2013*). This feature is associated with greater genetic variability and the generation of hypertransmittable subtypes and favors a wide distribution of gp60-based allelic families in the phylogenetic tree without a cluster aggregation. The recombination phenomenon was not a relevant variable in our study since genomes analyzed correspond to isolates circulating in different geographic locations of four continents.

Several authors have proposed that phenotypic differences between *Cryptosporidium* species are related to polymorphisms on protein-coding regions (*Xu et al., 2004*; *Pain, Crossman & Parkhill, 2005*; *Bouzid et al., 2013*; *Nader et al., 2019*). In our study, the single nucleotide variants located in CDSs were analyzed and characterized as synonymous and non-synonymous changes. Compared to the *C. parvum* reference, *Cryptosporidium* species that showed the highest number of SNVs in protein-coding regions were *C. meleagridis* and *C. hominis* than 400,000 and 150,000 variants, respectively. Forty-two percent of them corresponded to non-synonymous changes. Our results on the number of SNVs in *C. hominis* differs from the data reported previously, and similarly, the same occurs with SNVs in coding regions. *Isaza et al. (2015)* identified 36,753 SNVs located on coding regions in the genome of four isolates of *C. hominis* (version 8 of CryptoDB), using as a reference the genome of *C. parvum* Iowa II. As we mentioned before, these discrepancies may be related to the methodological approach used in every study.

In *C. parvum*, the number of variants in coding regions was below 20,000 SNVs, with intraspecies differences related to a heterogeneous behavior between zoonotic and anthroponotic isolates. Additionally, it was identified that more than 50% of the SNVs located in coding regions correspond to non-synonymous mutations. In a previous study carried out by *Widmer et al. (2012)*, non-synonymous SNVs were present in a range from 28% to 32% of all SNVs, and 60% of all nucleotide positions in the two genomes were not synonymous.

Analysis of the deleterious mutations in the *Cryptosporidium* species evaluated in the study showed that 183 genes were predicted with mutations that were present in at least one genome of each species: 377 exclusive to *C. meleagridis*, 103 to *C. hominis,* and 33 to *C. parvum*. Unfortunately, 53,5% of the genes with deleterious changes were located in uncharacterized coding regions (hypothetical proteins), so the biological impact of these mutations could not be determined. Annotation of the shared and species-exclusive genes with deleterious mutations establish that most encode enzymes and proteins involved in DNA repair, recombination processes, proteins associated with the chromosomes, as well as the biogenesis of transfer RNA and ribosomes. Comparative analyzes carried out previously have found that the genes with the highest number of SNVs in *C. hominis* and

C. parvum were related to ribosome assembly, translation processes, and coding genes for proteins with transmembrane domains (Widmer et al., 2012; Isaza et al., 2015). Sikora et al. (2017b) carried out an intraspecies comparative analysis of 14 genomes of *C. hominis*. They found 18 genes with non-synonymous mutations, with only the gene that codes for a protein of the oocyst wall COWP9 (oocyst of cgd6_210) with a deleterious mutation, which was present in eight of the fourteen genomes. Mutated genes annotated in our study also code for surface proteins with transmembrane domains and proteins secreted by non-classical pathways, suggesting that the interaction processes between the parasite and the host cells could be affected. This finding agrees with that reported by other authors who have described that the processes that are mainly affected are adhesion and invasion (Li et al., 2013; Isaza et al., 2015; Gilchrist et al., 2018; Widmer, 2018; Xu et al., 2019). It has been determined that the proteins secreted by non-classical pathways usually are growth factors, inflammatory cytokines, components of the extracellular matrix that regulate cell differentiation, proliferation, and apoptosis, as well as surface proteins in parasites involved in the initial interaction with the host (Nickel, 2003). This reinforces the need to improve the annotation of *Cryptosporidium* genomes, allowing the understanding of unknown aspects related to evolution, virulence, and pathogenicity in this genus.

Indels events in the genomes of the 23 *Cryptosporidium* isolates were also evaluated in our study. More than 60% of these variants were located in CDSs, a finding that could be expected since the *Cryptosporidium* genome has a percentage of coding regions more significant than 70% (Nader et al., 2019; Xu et al., 2019). The highest number of deletions, using reference *C. parvum* IOWA, occurred in *C. meleagridis* genomes, followed by *C. hominis*, suggesting a partial loss of genome fragments these species. Indel events were less abundant than SNVs in the 23 genomes analyzed, contrary to the reports in other apicomplexan such as *Plasmodium falciparum*, in which they have been described as the dominant mechanism of polymorphism within the genome (Miles et al., 2016). Feng et al. (2017) identified 1,200 insertion events and 1,500 deletions in a comparative genomic study of *C. parvum* isolates. In our study, which analyzed a more extensive number of genomes, we found less than 100 structural variations in each of the zoonotic and anthroponotic isolates. Previous studies have determined that indels are significantly more frequent in the peri-telomeric and subtelomeric regions of *Cryptosporidium* genomes (Nader et al., 2019). Guo et al. (2015) analyzed five genomes of *C. hominis* against the reference *C. parvum* Iowa, and they identified several insertions and deletions near the telomeres on chromosome 6, associated with recombination events, which could indicate that the duplication or deletion of subtelomeric genes is involved in the differences in host specificity between *Cryptosporidium* species. These recombination events can also explain the low support obtained in the phylogeny for the *C. hominis* 30976 isolate. Members of multicopy gene families and under a strong positive selection, such as MEDLE proteins, insulin-like proteases, and mucin-type glycoproteins, related to the parasite-host interaction, are ubicated in these regions (Fan, Feng & Xiao, 2019; Feng & Xiao, 2019; Xu et al., 2019).

Although several studies have described deletions in genes encoding the MEDLE proteins in *C. parvum* and *C. hominis* (Widmer et al., 2012; Liu et al., 2016), in the present study, no structural variants were found in these genes or those encoding insulin-like proteases.

However, deletions were found in all genomes of *C. meleagridis* and at least two isolates of *C. hominis* and *C. parvum* in a gene coding for a cryptosporidial mucin (ortholog of cgd7_4020), also known as gp900. This is a microneme secreted surface glycoprotein encoded by a single copy gene; it is involved in the apical portion of sporozoites and merozoites to enterocytes, which is required to initiate the invasion process (*Okhuysen & Chappell, 2002*; *Carruthers & Tomley, 2008*). However, since there is a repertoire of adhesion proteins in *Cryptosporidium*, including the proteins related to thrombospondin, p23, the gp40 / p30 protein complex, and the Circumsporozoite-like protein—CSL (*Langer-Curry & Riggs, 1999*; *Bouzid et al., 2013*), the alterations identified in the gp900 gene probably do not affect the interaction processes between the invasive stages of the parasite and the host cell. Additional analyzes are required to determine the biological implications of these deletions in the binding and invasion process, mainly in isolates of *C. meleagridis*.

Another interesting finding in this study was identifying deletions in all genomes of *C. meleagridis* and *C. hominis* for genes that encode for proteins with a WD-40 (cgd6_4290 ortholog) and SNF2/DEXDc/HELICc domains (cdg7_420 ortholog). WD-40 domain has tryptophan-aspartic acid (WD) repeats of approximately 40 amino acids and is considered one of the ten most abundant protein domains in eukaryotes (*Xu & Min, 2011*; *Jain & Pandey, 2018*). The proteins that contain these repeats are involved in various cellular processes, acting as an adaptor in many different protein complexes or protein-DNA complexes, signal transduction, transcription, cell cycle regulation, and apoptosis; however, no enzymatic activity has been assigned (*Stirnimann et al., 2010*; *Xu & Min, 2011*). These domains have been reported as highly polymorphic in other apicomplexan protozoa, such as *Plasmodium falciparum*, suggesting the participation of WD-40 in basic cellular and metabolic processes (*Chahar et al., 2015*). SWI2/SNF2 (Switching defective -SWI and Sucrose nonfermenting-SNF) protein family are ATP-dependent chromatin remodeling factors that modulate the access of transcription factors to regulatory regions of genes (*Sullivan et al., 2013*). Previous reports indicate homologs of these domains in different apicomplexan, including *Plasmodium falciparum* (*Ji & Arnot, 1997*) and *Toxoplasma gondii* (*Sullivan et al., 2003*). *Cryptosporidium* parvum has 14 chromatin-remodeling SNF2/SWI2 ATPases (*Templeton et al., 2004*). Alterations in the genes encoding these proteins could affect the epigenetic regulatory mechanisms in this genus.

Exclusive insertions were identified in 50% of the *C. parvum* isolates in the cgd3_190 gene. This gene encodes a microneme secreted protein with epidermal growth factor - EGF domains like, which are involved in cell signaling (*Carruthers & Tomley, 2008*). In other apicomplexans, such as *Toxoplasma gondii*, these domains have been associated with adhesion processes to the host cell (*Huynh, Boulanger & Carruthers, 2014*). It has also been shown that in the presence of calcium, the EGF domains adopt an extended structure resistant to proteases, favoring the interaction of the N-terminal portion of the molecule with the host cell ligands, favoring invasion (*Carruthers & Tomley, 2008*).

CONCLUSIONS

Here we present, to our knowledge, the most comprehensive phylogenomic and genomic comparative analysis performed in the most relevant human infecting *Cryptosporidium* species, which includes complete genomes from different isolates, allelic families, and subtypes. Comparative analysis of more than 800,000 single nucleotide variable positions detected in 24 genomes of the three main species infecting humans (*C. parvum, C. hominis,* and *C. meleagridis*) generate a more robust analysis on the phylogenetic relationships between the *Cryptosporidium* species of human public health concern. This phylogenomic analysis also confirmed the *gp60* loci segregation pattern observed in subtype families. Most of the SNVs and indels detected in the study genomes were ubicated in coding regions. Genes with deleterious changes and indels were identified and annotated, whenever possible, in the three species. These mutated genes were associated with the processing of genetic information and enzymatic and metabolic processes; however, most of them remain uncharacterized and encode hypothetical proteins.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was funded by the COLCIENCIAS Convocatoria 777-2017 Para Proyectos de Ciencia, Tecnología e Innovación en Salud 2017, project name: "Estudio de prevalencia, diversidad genética y genoḿica de Cryptosporidium en población VIH positiva de Antioquia" code 115-777 57608, and Vicerrectoría de Investigación (CODI), Universidad de Antioquia, grant "Finalización del genoma de referencia de C. hominis aislado UdeA01", code 2017-16-171. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
COLCIENCIAS Convocatoria 777-2017 Para Proyectos de Ciencia, Tecnología e Innovación en Salud 2017, project name: "Estudio de prevalencia, diversidad genética y genoḿica de Cryptosporidium en población VIH positiva de Antioquia": 115-777 57608.
Vicerrectoría de Investigación (CODI), Universidad de Antioquia, grant "Finalización del genoma de referencia de C. hominis aislado UdeA01": 2017-16-171.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Laura M. Arias-Agudelo performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Gisela Garcia-Montoya and Ana L Galvan-Diaz performed the experiments, authored or reviewed drafts of the paper, and approved the final draft.
- Felipe Cabarcas performed the experiments, analyzed the data, authored or reviewed drafts of the paper, and approved the final draft.
- Juan F. Alzate conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The Cryptdb Assemblies datasets, ChominisUKH1, Chominis37999, ChominisTU502, Chominis30976, ChominisUdeA01, and CmeleagridisUKMEL1, are available at CryptoDB (Release-43 Download).

SRA NGS reads

Isolate ID	BioProject ID	SRA
UKH1	PRJNA222837	SRR1015721
UKH3	PRJNA253834	SRR6131684
UKH4	PRJNA253838	SRR6143718
UKH5	PRJNA253839	SRR6144056
37999	PRJNA252787	SRR1558150
TU502_2012	PRJNA222836	SRR1015747
30976	PRJNA252787	SRR1557959
UdeA01	PRJEB10000	ERX1047563
SWEH2	PRJNA307563	SRR3098103
SWEH5	PRJNA307563	SRR3098109
UKP2	PRJNA253836	SRR6117460
UKP3	PRJNA253840	SRR6147472
UKP4	PRJNA253843	SRR6147581
UKP5	PRJNA253845	SRR6147587
UKP6	PRJNA253846	SRR6147945
UKP7	PRJNA253847	SRR6147964
UKP8	PRJNA253848	SRR6148259
UKP14	PRJNA315506	SRR6813718
UKP15	PRJNA315507	SRR6813719
UKMEL1	PRJNA222838	SRR1179185
UKMEL3	PRJNA315502	SRR6813720
UKMEL4	PRJNA315503	SRR6813896
TU1867	PRJNA192428	SRR793561

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.10478#supplemental-information.

REFERENCES

Abal-Fabeiro JL, Maside X, Bello X, Llovo J, Bartolomé C. 2013. Multilocus patterns of genetic variation across Cryptosporidium species suggest balancing selection at the gp60 locus. *Molecular Ecology* **22**:4723–4732 DOI 10.1111/mec.12425.

Abe N, Makino I. 2010. Multilocus genotypic analysis of Cryptosporidium isolates from cockatiels, Japan. *Parasitology Research* **106**:1491–1497 DOI 10.1007/s00436-010-1810-5.

Abrahamsen MS. 2004. Complete genome sequence of the apicomplexan, cryptosporidium parvum. *Science* **304**:441–445 DOI 10.1126/science.1094786.

Adamu H, Petros B, Zhang G, Kassa H, Amer S, Ye J, Feng Y, Xiao L. 2014. Distribution and clinical manifestations of cryptosporidium species and subtypes in HIV/AIDS patients in Ethiopia. *PLOS Neglected Tropical Diseases* **8**:e2831 DOI 10.1371/journal.pntd.0002831.

Andrew R. 2018. FigTree, version 1.4.4. Computer program distributed by the author, website. *Available at http://tree.bio.ed.ac.uk/software/figtree/* accessed enero, 2020 (accessed on 15 October 2015).

Almagro Armenteros JJ, Tsirigos KD, Sønderby CK, Petersen TN, Winther O, Brunak S, Von Heijne G, Nielsen H. 2019. SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nature Biotechnology* **37**:420–423 DOI 10.1038/s41587-019-0036.

Baele G, Suchard MA, Rambaut A, Lemey P. 2017. Emerging concepts of data integration in pathogen phylodynamics. *Systematic Biology* **66**:e47–e65 DOI 10.1093/sysbio/syw054.

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham SON, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MAXA, Pevzner PA. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology* **19**:455–477 DOI 10.1089/cmb.2012.0021.

Bendtsen JD, Jensen LJ, Blom N, Von Heijne G, Brunak S. 2004. Feature-based prediction of non-classical and leaderless protein secretion. *Protein Engineering, Design and Selection* **17**:349–356 DOI 10.1093/protein/gzh037.

Beser J, Hallström BM, Advani A, Andersson S, Östlund G, Winiecka-Krusnell J, Lebbad M, Alm E, Troell K, Arrighi RBG. 2017. Improving the genotyping resolution of Cryptosporidium hominis subtype IbA10G2 using one step PCR-based amplicon sequencing. *Infection, Genetics and Evolution* **55**:297–304 DOI 10.1016/j.meegid.2017.08.035.

Blake DP, Clark EL, Macdonald SE, Thenmozhi V, Kundu K, Garg R, Jatau ID, Ayoade S, Kawahara F, Moftah A, Reid AJ, Adebambo AO, Zapata RÁ, Rao ASRS, Thangaraj K, Banerjee PS, Dhinakar-Raj G, Raman M, Tomley FM. 2015. Population, genetic, and antigenic diversity of the apicomplexan Eimeria tenella and their relevance to vaccine development. *Proceedings of the National Academy of Sciences of the United States of America* **112**:E5343–E5350 DOI 10.1073/pnas.1506468112.

Bouzid M, Hunter PR, Chalmers RM, Tyler KM. 2013. Cryptosporidium pathogenicity and virulence. *Clinical Microbiology Reviews* **26**:115–134 DOI 10.1128/CMR.00076-12.

Caccio SM, Widmer G. 2014. Cryptosporidium: parasite and disease. Vienna: Springer.

Camaa VA, Ross JMM, Crawford S, Kawai V, Chavez-Valdez R, Vargas D, Vivar A, Ticona E, Navincopa M, Williamson J, Ortega Y, Gilman RHH, Bern C, Xiao L, Chavez-Valdez R, Vargas D, Vivar A, Ticona E, Ñavincopa M, Williamson J, Ortega Y, Gilman RHH, Bern C, Xiao L. 2007. Differences in clinical manifestations among cryptosporidium species and subtypes in HIV-infected persons. *The Journal of Infectious Diseases* **196**:684–691 DOI 10.1086/519842.

Cama VA, Bern C, Roberts J, Cabrera L, Sterling CR, Ortega Y, Gilman RH, Xiao L. 2008. Cryptosporidium species and subtypes and clinical manifestations in children, Peru. *Emerging Infectious Diseases* **14**:1567–1574 DOI 10.3201/eid1410.071273.

Carruthers VB, Tomley FM. 2008. Microneme proteins in apicomplexans. *Sub-Cellular Biochemistry* **47**:33–45 DOI 10.1007/978-0-387-78267-6_2.

Chahar P, Kaushik M, Gill SS, Gakhar SK, Gopalan N, Datt M, Sharma A, Gill R. 2015. Genome-wide collation of the Plasmodium falciparum WDR protein superfamily reveals malarial parasite-specific features. *PLOS ONE* **10**:1–24 DOI 10.1371/journal.pone.0128507.

Cornillot E, Hadj-Kaddour K, Dassouli A, Noel B, Ranwez V, Vacherie B, Augagneur Y, Brès V, Duclos A, Randazzo S, Carcy B, Debierre-Grockiego F, Delbecq S, Moubri-Ménage K, Shams-Eldin H, Usmani-Brown S, Bringaud F, Wincker P, Vivarès CP, Schwarz RT, Schetters TP, Krause PJ, Gorenflot A, Berry V, Barbe V, Mamoun Ben C. 2012. Sequencing of the smallest Apicomplexan genome from the human pathogen Babesia microti. *Nucleic Acids Research* **40**:9102–9114 DOI 10.1093/nar/gks700.

Cunha FS, Peralta JM, Peralta RHS. 2019. New insights into the detection and molecular characterization of cryptosporidium with emphasis in Brazilian studies: a review. *Revista do Instituto de Medicina Tropical de Sao Paulo* **61**: DOI 10.1590/s1678-9946201961028.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R. 2011. The variant call format and VCFtools. *Bioinformatics* **27**:2156–2158 DOI 10.1093/bioinformatics/btr330.

Danišová O, Valenčáková A, Stanko M, Luptáková L, Hatalová E, Čanády A. 2017. Rodents as a reservoir of infection caused by multiple zoonotic species/genotypes of C. parvum, C. hominis, C. suis, C. scrofarum, and the first evidence of C. muskrat genotypes I and II of rodents in Europe. *Acta Tropica* **172**:29–35 DOI 10.1016/j.actatropica.2017.04.013.

Delcher AL, Phillippy A, Carlton J, Salzberg SL. 2002. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Research* **30(11)**:2478–2483 DOI 10.1093/nar/30.11.2478.

DeMaio N, Shaw LP, Hubbard A, George S, Sanderson ND, Swann J, Wick R, Oun MA, Stubberfield E, Hoosdally SJ, Crook DW, Peto TEA, Sheppard AE, Bailey MJ, Read DS, Anjum MF, Sarah Walker A, Stoesser N. 2019. Comparison of long-read sequencing technologies in the hybrid assembly of complex bacterial genomes. *Microbial Genomics* **5**:1–21 DOI 10.1099/mgen.0.000294.

Efstratiou A, Ongerth JE, Karanis P. 2017. Waterborne transmission of protozoan parasites: review of worldwide outbreaks - an update 2011–2016. *Water Research* **114**:14–22 DOI 10.1016/j.watres.2017.01.036.

Fan Y, Feng Y, Xiao L. 2019. Comparative genomics: how has it advanced our knowledge of cryptosporidiosis epidemiology? *Parasitology Research* **118**:3195–3204 DOI 10.1007/s00436-019-06537-x.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution* **17**:368–376 DOI 10.1007/BF01734359.

Feng Y, Alderisio KA, Yang W, Blancero LA, Kuhne WG, Nadareski CA, Reid M, Xiao L. 2007. Cryptosporidium genotypes in wildlife from a New York watershed. *Applied and Environmental Microbiology* **73**:6475–6483 DOI 10.1128/AEM.01034-07.

Feng Y, Li N, Roellig DM, Kelley A, Liu G, Amer S, Tang K, Zhang L, Xiao L. 2017. Comparative genomic analysis of the IId subtype family of Cryptosporidium parvum. *International Journal for Parasitology* **47**:281–290 DOI 10.1016/j.ijpara.2016.12.002.

Feng Y, Ryan UM, Xiao L. 2018. Genetic diversity and population structure of cryptosporidium. *Trends in Parasitology* **34**:997–1011 DOI 10.1016/j.pt.2018.07.009.

Feng Y, Xiao L. 2019. Differential expression of three cryptosporidium species-specific MEDLE Proteins. *Frontiers in Microbiology* **10**:1–11 DOI 10.3389/fmicb.2019.01177.

Firoozi Z, Sazmand A, Zahedi A, Astani A, Fattahi-Bafghi A, Kiani-Salmi N, Ebrahimi B, Dehghani-Tafti A, Ryan U, Akrami-Mohajeri F. 2019. Prevalence and genotyping identification of Cryptosporidium in adult ruminants in central Iran. *Parasites and Vectors* **12**:510 DOI 10.1186/s13071-019-3759-2.

Galen SC, Borner J, Martinsen ES, Schaer J, Austin CC, West CJ, Perkins SL. 2018. The polyphyly of Plasmodium: comprehensive phylogenetic analyses of the malaria parasites (Order Haemosporida) reveal widespread taxonomic conflict. *Royal Society Open Science* **5(5)**:171780 DOI 10.1098/rsos.171780.

Gilchrist CA, Cotton JA, Burkey C, Arju T, Gilmartin A, Lin Y, Ahmed E, Steiner K, Alam M, Ahmed S, Robinson G, Zaman SU, Kabir M, Sanders M, Chalmers RM, Ahmed T, Ma JZ, Haque R, Faruque ASG, Berriman M, Petri WA. 2018. Genetic diversity of cryptosporidium hominis in a bangladeshi community as revealed by whole-genome sequencing. *Journal of Infectious Diseases* **218**:259–264 DOI 10.1093/infdis/jiy121.

Glaberman S, Sulaiman IM, Bern C, Limor J, Peng MM, Morgan U, Gilman R, Lal AA, Xiao L. 2001. A multilocus genotypic analysis of Cryptosporidium meleagridis. *Journal of Eukaryotic Microbiology* **Suppl**:9S–22S DOI 10.1111/j.1550-7408.2001.tb00439.x.

Grinberg A, Widmer G. 2016. Cryptosporidium within-host genetic diversity: systematic bibliographical search and narrative overview. *International Journal for Parasitology* **46**:465–471 DOI 10.1016/j.ijpara.2016.03.002.

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology* **59**:307–321 DOI 10.1093/sysbio/syq010.

Guo Y, Tang K, Rowe LA, Li N, Roellig DM, Knipe K, Frace M, Yang C, Feng Y, Xiao L. 2015. Comparative genomic analysis reveals occurrence of genetic recombination in virulent Cryptosporidium hominis subtypes and telomeric gene duplications in Cryptosporidium parvum. *BMC Genomics* **16**:320 DOI 10.1186/s12864-015-1517-1.

Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: improving the Ultrafast Bootstrap Approximation. *Molecular Biology and Evolution* **35**:518–522 DOI 10.1093/molbev/msx281.

Holder A, Simon J, Strauser J, Taylor J, Shibberu Y. 2013. Dynamic programming used to align protein structures with a spectrum is robust. *Biology* **2**:1296–1310 DOI 10.3390/biology2041296.

Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, Von Mering C, Bork P. 2017. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Molecular Biology and Evolution* **34**:2115–2122 DOI 10.1093/molbev/msx148.

Huynh MH, Boulanger MJ, Carruthers VB. 2014. A conserved apicomplexan microneme protein contributes to Toxoplasma gondii invasion and virulence. *Infection and Immunity* **82**:4358–4368 DOI 10.1128/IAI.01877-14.

Ifeonu OO, Chibucos MC, Orvis J, Su Q, Elwin K, Guo F, Zhang H, Xiao L, Sun M, Chalmers RM, Fraser CM, Zhu G, Kissinger JC, Widmerg G, Silva JC. 2016. Annotated draft genome sequences of three species of Cryptosporidium: C. meleagridis isolate UKMEL1, C. baileyi isolate TAMU-09Q1, and C. hominis isolates TU502_2012 and UKH1. *Pathogens and Disease* **74**:ftw080 DOI 10.1093/femspd/ftw080.

Isaza JP, Galvan AL, Polanco V, Huang B, Matveyev AV, Serrano MG, Manque P, Buck GA, Alzate JF. 2015. Revisiting the reference genomes of human pathogenic Cryptosporidium species: reannotation of C. parvum Iowa and a new C. hominis reference. *Scientific Reports* **5**:16324 DOI 10.1038/srep16324.

Jain BP, Pandey S. 2018. WD40 repeat proteins: signalling scaffold with diverse functions. *Protein Journal* **37**:391–406 DOI 10.1007/s10930-018-9785-7.

Ji D-D, Arnot DE. 1997. A Plasmodium falciparum homologue of the ATPase subunit of a multi- protein complex involved in chromatin remodelling for transcription. *Molecular and Biochemical Parasitology* **88(97)**:151–162 DOI 10.1016/S0166-6851(97)00089-3.

Khan A, Shaik JS, Grigg ME. 2017. Genomics and molecular epidemiology of Cryptosporidium species. *Acta Tropica* **184**:1–14 DOI 10.1016/j.actatropica.2017.10.023.

Krogh A, Larsson B, Von Heijne G, Sonnhammer ELL. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of Molecular Biology* **305**:567–580 DOI 10.1006/jmbi.2000.4315.

Lack JB, Reichard MV, Van Den Bussche RA. 2012. Phylogeny and evolution of the Piroplasmida as inferred from 18S rRNA sequences. *International Journal for Parasitology* **42**:353–363 DOI 10.1016/j.ijpara.2012.02.005.

Langer-Curry R, Riggs MW. 1999. Cryptosporidium parvum apical complex glycoprotein CSL contains a sporozoite ligand for intestinal epithelial cells. *Infection and Immunity* **67**:5282–5291 DOI 10.1128/IAI.67.10.5282-5291.1999.

Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**:589–595 DOI 10.1093/bioinformatics/btp698.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**:2078–2079 DOI 10.1093/bioinformatics/btp352.

Li N, Xiao L, Cama VA, Ortega Y, Gilman RH, Guo M, Feng Y. 2013. Genetic recombination and Cryptosporidium hominis virulent subtype IbA10G2. *Emerging Infectious Diseases* **19**:1573–1582 DOI 10.3201/eid1910.121361.

Liu S, Roellig DM, Guo Y, Li N, Frace MA, Tang K, Zhang L, Feng Y, Xiao L. 2016. Evolution of mitosome metabolism and invasion-related proteins in Cryptosporidium. *BMC Genomics* **17**:1–16 DOI 10.1186/s12864-016-3343-5.

Magoc T, Salzberg SL. 2011. FLASH : fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**:2957–2963 DOI 10.1093/bioinformatics/btr507.

Mazurie AJ, Alves JM, Ozaki LS, Zhou S, Schwartz DC, Buck GA. 2013a. Comparative genomics of cryptosporidium. *International Journal of Genomics* **2013**:832756 DOI 10.1155/2013/832756.

Mazurie AJ, Alves JM, Ozaki LS, Zhou S, Schwartz DC, Buck GA. 2013b. Comparative genomics of Cryptosporidium. *Hindawi Publishing Corporation* **2013**:1–9 DOI 10.1155/2013/832756.

Miles A, Iqbal Z, Vauterin P, Pearson R, Campino S, Theron M, Gould K, Mead D, Drury E, O'Brien J, Rubio VR, Macinnis B, Mwangi J, Samarakoon U, Ranford-Cartwright L, Ferdig M, Hayton K, Su XZ, Wellems T, Rayner J, McVean G, Kwiatkowski D. 2016. Indels, structural variation, and recombination drive genomic diversity in Plasmodium falciparum. *Genome Research* **26**:1288–1299 DOI 10.1101/gr.203711.115.

Morris A, Robinson G, Swain MT, Chalmers RM. 2019. Direct sequencing of cryptosporidium in stool samples for public health. *Frontiers in Public Health* **7**:360 DOI 10.3389/fpubh.2019.00360.

Nader JL, Mathers TC, Ward BJ, Pachebat JA, Swain MT, Robinson G, Chalmers RM, Hunter PR, van Oosterhout C, Tyler KM. 2019. Evolutionary genomics of anthroponosis in Cryptosporidium. *Nature Microbiology* **4**:826–836 DOI 10.1038/s41564-019-0377.

Nattestad M, Schatz MC. 2016. Assemblytics: A web analytics tool for the detection of variants from an assembly. *Bioinformatics* **32**:3021–3023 DOI 10.1093/bioinformatics/btw369.

Nickel W. 2003. The mystery of nonclassical protein secretion: a current view on cargo proteins and potential export routes. *European Journal of Biochemistry* **270**:2109–2119 DOI 10.1046/j.1432-1033.2003.03577.x.

Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. 1999. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research* **27**:29–34 DOI 10.1093/nar/27.1.29.

Okhuysen PC, Chappell CL. 2002. Cryptosporidium virulence determinants - are we there yet? *International Journal for Parasitology* **32**:517–525 DOI 10.1016/S0020-7519(01)00356-3.

Pain A, Crossman L, Parkhill J. 2005. Comparative Apicomplexan genomics. *Nature* **3**:454–455 DOI 10.1038/nrmicro1164.

Pérez-Cordón G, Robinson G, Nader J, Chalmers RM. 2016. Discovery of new variable number tandem repeat loci in multiple Cryptosporidium parvum genomes for the surveillance and investigation of outbreaks of cryptosporidiosis. *Experimental Parasitology* **169**:119–128 DOI 10.1016/j.exppara.2016.08.003.

Puiu D. 2004. CryptoDB: the Cryptosporidium genome resource. *Nucleic Acids Research* **32(329D-331)**: DOI 10.1093/nar/gkh050.

Ren X, Zhao J, Zhang L, Ning C, Jian F, Wang R, Lv C, Wang Q, Arrowood MJ, Xiao L. 2012. Cryptosporidium tyzzeri n. sp. (Apicomplexa: Cryptosporididae) in domestic mice (Mus musculus). *Experimental Parasitology* **130**:274–81 DOI 10.1016/j.exppara.2011.07.012.

Ryan U, Fayer R, Xiao L. 2014. Cryptosporidium species in humans and animals: current understanding and research needs. *Parasitology* **141**:1667–85 DOI 10.1017/S0031182014001085.

Ryan U, Paparini A, Monis P, Hijjawi N. 2016. It's official- Cryptosporidium is a gregarine: what are the implications for the water industry? *Water Research* **105**:305–313 DOI 10.1016/j.watres.2016.09.013.

Schwarz G. 1978. Estimating the dimension of a model. *The Annals of Statistics* **6**:461–464 DOI 10.1214/aos/1176344136.

Shimodaira H, Hasegawa M. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Molecular Biology and Evolution* **16**:1114–1116 DOI 10.1093/oxfordjournals.molbev.a026201.

Sikora P, Andersson S, Winiecka-Krusnell J, Hallström B, Alsmark C, Troell K, Beser J, Arrighi RBG. 2017a. Genomic variation in IbA10G2 and other patient-derived cryptosporidium hominis subtypes. *Journal of Clinical Microbiology* **55**:844–858 DOI 10.1128/JCM.01798-16.

Sikora P, Arrighi RBG, Beser J, Andersson S. 2017b. Genomic variation in IbA10G2 and other patient-derived cryptosporidium hominis subtypes. *Journal of Clinical Microbiology* **55**:844–858 DOI 10.1128/JCM.01798-16.

Sow SO, Muhsen K, Nasrin D, Blackwelder WC, Wu Y, Farag TH, Panchalingam S, Sur D, Zaidi AKM, Faruque ASG, Saha D, Adegbola R, Alonso PL, Breiman RF, Bassat Q, Tamboura B, Sanogo D, Onwuchekwa U, Manna B, Ramamurthy T, Kanungo S, Ahmed S, Qureshi S, Quadri F, Hossain A, Das SK, Antonio M, Hossain MJ, Mandomando I, Nhampossa T, Acácio S, Omore R, Oundo JO, Ochieng JB, Mintz ED, O'Reilly CE, Berkeley LY, Livio S, Tennant SM, Sommerfelt H, Nataro JP, Ziv-Baran T, Robins-Browne RM, Mishcherkin V, Zhang J, Liu J, Houpt ER, Kotloff KL, Levine MM. 2016. The Burden of Cryptosporidium Diarrheal Disease among Children <24 months of age in moderate/high mortality regions of sub-saharan africa and south asia, utilizing data from the global enteric multicenter study (GEMS). *PLOS Neglected Tropical Diseases* **10**:e0004729 DOI 10.1371/journal.pntd.0004729.

Stensvold CR, Beser J, Axen C, Lebbad M. 2014. High applicability of a novel method for gp60-based subtyping of Cryptosporidium meleagridis. *Journal of Clinical Microbiology* **52**:2311–2319 DOI 10.1128/JCM.00598-14.

Stirnimann CU, Petsalaki E, Russell RB, Christoph WM. 2010. WD40 proteins propel cellular networks. *Review Trends in Biochemical Sciences* **35**:531–538 DOI 10.1016/j.tibs.2010.04.003.

Su J, Jin C, Wu H, Fei J, Li N, Guo Y, Feng Y, Xiao L. 2019. Differential expression of three cryptosporidium species-specific MEDLE proteins. *Frontiers in Microbiology* **10**:1177 DOI 10.3389/fmicb.2019.01177.

Sulaiman IM, Lal AA, Xiao L. 2002. Molecular phylogeny and evolutionary relationships of cryptosporidium parasites at the actin locus. *The Journal of Parasitology* **88**(**2**):388–394 DOI 10.2307/3285594.

Sullivan WJ, Monroy MA, Bohne W, Nallani KC, Chrivia J, Yaciuk P, Smith CK, Queener SF. 2003. Molecular cloning and characterization of an SRCAP chromatin remodeling homologue in Toxoplasma gondii. *Parasitology Research* **90**:1–8 DOI 10.1007/s00436-002-0814-1.

Sullivan WJ, Radke JB, Kim K, White MW. 2013. *Epigenetic and genetic factors that regulate gene expression in toxoplasma gondii.* Cambridge: Academic Press DOI 10.1016/B978-0-12-396481-6.00018-0.

Templeton TJ, Iyer LM, Anantharaman V, Enomoto S, Abrahante JE, Subramanian GM, Hoffman SL, Abrahamsen MS, Aravind L. 2004. Comparative analysis of apicomplexa and genomic diversity in eukaryotes. *Genome Research* **14**:1686–1695 DOI 10.1101/gr.2615304.

Theys K, Lemey P, Vandamme AM, Baele G. 2019. Advances in visualization tools for phylogenomic and phylodynamic studies of viral diseases. *Frontiers in Public Health* **7**:208 DOI 10.3389/fpubh.2019.00208.

Trifinopoulos J, Nguyen LT, von Haeseler A, Minh BQ. 2016. W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. *Nucleic Acids Research* **44**:W232–W235 DOI 10.1093/nar/gkw256.

Troeger C, Forouzanfar M, Rao PC, Khalil I, Brown A, Reiner RC, Fullman N, Thompson RL, Abajobir A, Ahmed M, Alemayohu MA, Alvis-Guzman N, Amare AT, Antonio CA, Asayesh H, Avokpaho E, Awasthi A, Bacha U, Barac A, Betsue BD, Beyene AS, Boneya DJ, Malta DC, Dandona L, Dandona R, Dubey M, Eshrati B, Fitchett JRA, Gebrehiwot TT, Hailu GB, Horino M, Hotez PJ, Jibat T, Jonas JB, Kasaeian A, Kissoon N, Kotloff K, Koyanagi A, Kumar GA, Rai RK, Lal A, Razek HMAEl, Mengistie MA, Moe C, Patton G, Platts-Mills JA, Qorbani M, Ram U, Roba HS, Sanabria J, Sartorius B, Sawhney M, Shigematsu M, Sreeramareddy C, Swaminathan S, Tedla BA, Jagiellonian RTM, Ukwaja K, Werdecker A, Widdowson MA, Yonemoto N, Zaki MElSayed, Lim SS, Naghavi M, Vos T, Hay SI, Murray CJL, Mokdad AH. 2017. Estimates of global,regional, and national morbidity, mortality, and aetiologies of diarrhoeal diseases: a systematic analysis for the Global Burden of Disease Study 2015. *The Lancet Infectious Diseases* **17**:909–948 DOI 10.1016/S1473-3099(17)30276-1.

Vaser R, Adusumalli S, Leng SN, Sikic M, Ng P. 2016. SIFT missense predictions for genomes. *Nature Protocols* **11**:1073–1081 DOI 10.1038/nprot.2015.123.

Šlapeta J. 2013. Cryptosporidiosis and Cryptosporidium species in animals and humans: a thirty colour rainbow? *International Journal for Parasitology* **43**:957–970 DOI 10.1016/j.ijpara.2013.07.005.

Wagnerová P, Sak B, McEvoy J, Rost M, Perec Matysiak A, Ježková J, Kváč M. 2015. Genetic diversity of Cryptosporidium spp. including novel identification of the Cryptosporidium muris and Cryptosporidium tyzzeri in horses in the Czech Republic and Poland. *Parasitology Research* **114**:1619–1624 DOI 10.1007/s00436-015-4353.

Weir W, Karagenç T, Gharbi M, Simuunza M, Aypak S, Aysul N, Darghouth MA, Shiels B, Tait A. 2011. Population diversity and multiplicity of infection in Theileria annulata. *International Journal for Parasitology* **41**:193–203 DOI 10.1016/j.ijpara.2010.08.004.

Widmer G. 2018. Diverse single-amino-acid repeat profiles in the genus Cryptosporidium. *Mineralogical Magazine* **145**:1151–1160 DOI 10.1017/S0031182018000112.

Widmer G, Lee Y, Hunt P, Martinelli A, Tolkoff M, Bodi K. 2012. Comparative genome analysis of two Cryptosporidium parvum isolates with different host range. *Infection, Genetics and Evolution* **12**:1213–1221 DOI 10.1016/j.meegid.2012.03.027.

Widmer G, Sullivan S. 2012. Genomics and population biology of Cryptosporidium species. *Parasite Immunology* **34**:61–71 DOI 10.1111/j.1365-3024.2011.01301.

Xiao L. 2010. Molecular epidemiology of cryptosporidiosis: an update. *Experimental Parasitology* **124**:80–89 DOI 10.1016/j.exppara.2009.03.018.

Xiao L, Feng Y. 2017. Molecular epidemiologic tools for waterborne pathogens Cryptosporidium spp. and Giardia duodenalis. *Food and Waterborne Parasitology* **8–9**:14–32 DOI 10.1016/j.fawpar.2017.09.002.

Xu C, Min J. 2011. Structure and function of WD40 domain proteins. *Protein and Cell* **2**:202–214 DOI 10.1007/s13238-011-1018-1.

Xu P, Widmer G, Wang Y, Ozaki LS, Alves JM, Serrano MG, Puiu D, Manque P, Akiyoshi D, Mackey AJ, Pearson WR, Dear PH, Bankier AT, Peterson DL, Abrahamsen MS, Kapur V, Tzipori S, Buck GA. 2004. The genome of Cryptosporidium hominis. *Nature* **431**:1103–1107 DOI 10.1038/nature02990.

Xu Z, Guo Y, Roellig DM, Feng Y, Xiao L. 2019. Comparative analysis reveals conservation in genome organization among intestinal Cryptosporidium species and sequence divergence in potential secreted pathogenesis determinants among major human-infecting species. *BMC Genomics* **20**:406 DOI 10.1186/s12864-019-5788-9.

Ye J, Zhang Y, Cui H, Liu J, Wu Y, Cheng Y, Xu H, Huang X, Li S, Zhou A, Zhang X, Bolund L, Chen Q, Wang J, Yang H, Fang L, Shi C. 2018. WEGO 2.0: a web tool for analyzing and plotting GO annotations, 2018 update. *Nucleic Acids Research* **46**:W71–W75 DOI 10.1093/nar/gky400.

Young AD, Gillung JP. 2020. Phylogenomics —principles, opportunities and pitfalls of big-data phylogenetics. *Systematic Entomology* **45**:225–247 DOI 10.1111/syen.12406.

Zahedi A, Gofton AW, Jian F, Paparini A, Oskam C, Ball A, Robertson I, Ryan U, Blunt DS, Khramtsov NV, Upton SJ, Montelone BA, Isaza JP, Galvan AL, Polanco V, Huang B, Matveyev AV, Serrano MG, Manque P, Buck GA, Alzate JF, Of PB, Guo Y, Tang K, Rowe LA, Li N, Roellig DM, Knipe K, Frace MA, Yang C, Feng Y, Xiao L, Liu S, Roellig DM, Guo Y, Li N, Frace MA, Tang K, Zhang L, Feng Y, Xiao L, Bouzid M, Hunter PR, Chalmers RM, Tyler KM. 2013. Comparative genomic analysis reveals occurrence of genetic recombination in virulent Cryptosporidium hominis subtypes and telomeric gene duplications in Cryptosporidium parvum. *BMC Genomics* **5**:1–14 DOI 10.1186/s12864-015-1517-1.

Genetics and Genomics of Chronic Obstructive Pulmonary Disease

Edwin K. Silverman[1,2], **Avrum Spira**[3], **and Peter D. Paré**[4]

[1]Channing Laboratory, and [2]Division of Pulmonary and Critical Care Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts; [3]Division of Pulmonary and Critical Care Medicine, Translational Bioinformatics Program, Clinical and Translational Science Institute, Boston University, Boston, Massachusetts; and [4]James Hogg iCAPTURE Centre for Cardiovascular and Pulmonary Research and the Department of Medicine, St. Paul's Hospital, University of British Columbia, Vancouver, British Columbia, Canada

Chronic obstructive pulmonary disease (COPD) fulfills criteria for a complex genetic disease in which environmental factors interact with multiple polymorphic genes to influence susceptibility. Finding the genes that influence susceptibility can be approached in hypothesis testing or unbiased study designs. In candidate gene association studies, genetic variation in, and/or levels of, expression of genes known or suspected to be involved in the pathogenesis of COPD are compared in affected and unaffected individuals. Although this approach is useful it is limited by our present knowledge of disease pathophysiology. Genomewide studies of gene expression and of genetic variation are now possible and are not constrained by our limited knowledge. Although both of these unbiased approaches are in their infancy, they have already provided exciting new avenues for future investigation and potentially now approaches to risk prediction and therapy.

Keywords: chronic obstructive pulmonary disease; genetics; genomics

At this symposium, three talks were given related to the genetics and genomics of chronic obstructive pulmonary disease (COPD). Peter Paré concentrated on candidate gene studies in which biologically plausible genes have been tested for variants that may impart increased risk for a rapid decline of lung function among smokers. Edwin Silverman focused on the promising approach offered by genome-wide association (GWA) studies, and Avrum Spira reviewed the state of the art with respect to the examination of gene expression at the mRNA level and its relationship to COPD phenotype.

CANDIDATE GENE STUDIES

In candidate gene association studies, genetic variants in genes thought to be involved in pathobiological pathways leading to COPD are tested for statistical association with one or more COPD phenotype. Sandford and his colleagues have taken advantage of the well-characterized Lung Health Study cohort study to examine gene variants associated with rapid decline in lung function (1). They selected the phenotypic extremes from among the approximately 6,000 individuals in that study; the approximately 300 continued smokers who had the most rapid and the approximately 300 who had the least rapid decline in FEV_1 over the first 5 years of the study. To date they have examined over 50 genes, and Table 1 shows a summary of those in which they have found polymorphisms that impart increased or decreased risk for rapid decline. Also shown is the frequency of the risk or protective allele and the odds ratio associated with that allele. Finally, the attributable risk for the specific alleles associated with increased risk is shown. The attributable risk estimates the contribution of that polymorphism to the overall risk of being in the fast or slow decline group in this "population." The association of a number of these genes with rate of decline of lung function has been replicated in independent populations (2–7).

Although such candidate gene studies are useful, they are limited by their very nature to the study of genes that are known to have some function that could contribute to the development of, or protection from, COPD. It is highly likely that many genes with unknown function contribute to pathogenesis, but until recently it has not been practical to interrogate the entire genome either at the DNA or RNA level. Recent advances in genome-wide genotyping and transcriptomic analysis make such an unbiased approach possible.

GENOME-WIDE ASSOCIATION STUDIES

Technical improvements in SNP genotyping within the past 5 years have led to markedly increased genotyping capacity at reduced cost. Combined with the identification of millions of common SNPs through the HapMap project, high-throughput genotyping approaches have made GWA studies a feasible study design in complex disease genetics. Rather than beginning with family-based linkage studies or candidate gene studies to limit the genomic regions of interest, GWA studies include a panel of hundreds of thousands of SNPs across the entire genome (8). Although there are approximately 10 million common SNPs in the human genome, approximately 500,000 well-selected SNPs can adequately capture the information of most SNPs in white and Asian populations (9). Because hundreds of thousands of statistical tests of genetic association are performed, stringent thresholds to establish statistical significance are required in GWA studies. Nonetheless, GWA studies have unequivocally identified genome-wide significant associations in a rapidly growing number of complex diseases, including age-related macular degeneration, diabetes mellitus, Crohn's disease, and asthma (10).

In a recent collaborative GWA study led by Drs. Sreekumar Pillai and David Goldstein (11), 823 subjects with COPD and 810 smoking control subjects from Norway underwent genome-wide SNP genotyping with the Illumina HumanHap550 SNP panel. Association analysis for COPD affection status was performed using logistic regression analysis with covariates including age, sex, current smoking status, and pack-years of smoking. After GWA analysis was performed, the top 100 SNPs were genotyped in 1,891 individuals from 606 pedigrees in the International COPD Genetics Network (ICGN). Family-based association analysis for COPD was performed using the same covariates as in the Norway population, and seven SNPs with significant *P* values in both populations were tested for further replication in 389 subjects with COPD from the National Emphysema Treatment Trial (NETT) and 472 smoking control subjects from the Normative Aging Study

---

(*Received in original form April 26, 2009; accepted in final form June 30, 2009*)

Supported by NIH grant HL086936 (Characterization of COPD and Modulation of Disease Progression, to Jeanine D'Armiento)

Correspondence and requests for reprints should be addressed to Peter D. Paré, M.D., UBC James Hogg iCAPTURE Centre for Cardiovascular and Pulmonary Research, St. Paul's Hospital, 1081 Burrard Street, Vancouver, BC, V6Z 1Y6 Canada. E-mail: ppare@mrl.ubc.ca

Proc Am Thorac Soc Vol 6. pp 539–542, 2009
DOI: 10.1513/pats.200904-021DS
Internet address: www.atsjournals.org

TABLE 1. SIGNIFICANT GENETIC ASSOCIATIONS WITH RAPID DECLINE IN LUNG FUNCTION IN THE LUNG HEALTH STUDY

Candidate Gene	Odds Ratio	Relative Frequencies	Attributable Risk (%)
a1-AT Z	2.8 (1.2, 7.3)	6% vs. 3%	5
mEH Tyr113 > His, Arg139 > His *	2.4 (1.1–5.4)	7% vs. 3%	4
β2AR Glu27 > Gln	0.6 (0.4–0.8)	44% vs. 57%	
Haplotype IL1RN A1/IL1B –511T	1.4 (1–(2.1)	14% vs. 10%	4
Haplotype IL1RN A2/IL1B –511C	1.5 (1.1–2.1)	20% vs. 14%	7
IL1RN	1.6 (1.1–2.3)	72% vs. 62%	3
IL8 A-251T	1.32	58% vs. 51%	5
IL6	1.9 (1.1–3.2)	24% vs. 14%	11
MMP-1 -1607+G	1.6 (1.1–2.4)	28% vs. 19%	10
IL4RA	2.2	7% vs. 4%	5
Glutathione Transferases* (GSTP1 Ile105 > Val, GSTT1& GSTM1)	2.8 (1.1–7.2)	5% vs. 2%	3
Heme oxygenase-1	1.6 (1.0–2.3)	25% vs. 18%	10
Catalase	1.5 (1.1–2.2)	30% vs. 22%	10
Nicotine Receptor (CHRNA5)	3.0 (1.2–7.6)	41% vs. 36%	4

The candidate genes are: a1-ATZ = heterozygosity for the Z allele of alpha one antitrypsin; mEH = microsomal epoxide hydrolase; β2AR = the beta 2 adrenergic receptor; IL1RN A = the interleukin 1 receptor antagonist; IL8 = interleukin 8; IL6 = interleukin 6; MMP-1 = the matrix metalloproteinase 1; IL4R = the interleukin 4 receptor; glutathione transferases (GSTP1, GSTT1 and GSTM1); heme oxygenase-1; catalase; and the nicotine receptor (CHRNA5).

Other investigators have studied the same or additional genes in different cohorts of smokers for whom longitudinal lung function data are available. These studies confirmed the contributions of variants in microsomal epoxide hydrolase, heme oxygenase-1, and the glutathione S-transferases and have also implicated ADAM 33, cytochrome P450, and glutamate cystein ligase (2–7).

* For microsomal epoxide hydrolase and the glutathione transferases the odds ratios, frequencies, and attributable risk pertain to the presence of a compound genotype consisting of the presence of 2 (mEH) or 3 (glutathiones) risk alleles.

(NAS). Combined *P* values from all three study populations were calculated. A SNP on chromosome 15 (rs8034191) was identified with genome-wide significant association to COPD ($P = 2 \times 10^{-10}$), with the same direction of association (i.e., the same risk allele) in all three populations. There are multiple genes of interest near the most highly associated SNP, including several subunits of the nicotinic acetylcholine receptor (CHRNA3 and CHRNA5) and an iron-binding protein (IREB2). Because of the extensive linkage disequilibrium in this region, it is not possible to identify one key candidate gene; intensive study of all of the genes from this region will be required to identify all common variants and to determine if rare variants also contribute to COPD susceptibility. Of interest, this locus was also recently associated with lung cancer susceptibility in three independent reports (12–14), with peripheral vascular disease, and with nicotine addiction (15) (Table 2), suggesting that this locus may be associated with both COPD and lung cancer by influencing smoking behavior. A study from New Zealand has demonstrated associations of this genomic region with both COPD and lung cancer (16), raising the possibility that the lung cancer associations could be confounded by the frequent co-occurrence of COPD and lung cancer.

The second GWA study of a COPD-related phenotype was performed in 7,691 participants in the Framingham Heart Study (17). FEV_1/FVC was the primary phenotype analyzed. Four SNPs on chromosome 4q were strongly associated with FEV_1/FVC in Framingham, and one of these SNPs was also associated with FEV_1/FVC in the Family Heart Study. Of interest, SNPs in this same region also showed evidence for association with COPD in the Norway case-control study, ICGN, and NETT-NAS studies (11). Although FEV_1/FVC was the primary phenotype analyzed in the Framingham Heart Study and Family Heart Study, significant evidence for association to COPD affection status was also observed in the Family Heart Study. The most strongly associated SNP is near the gene that codes for the hedgehog interacting protein (HHIP), but the functional variant or variants in this region have not been identified. Significant associations to the CHRNA3/5 region were not observed in the Framingham population.

Although GWA studies in COPD are in their infancy, the chromosome 4 and 15 genetic associations are the most statistically compelling associations reported to date. Additional GWA studies of COPD and COPD-related pheno-

TABLE 2. GENOME-WIDE ASSOCIATION STUDIES WITH SIGNIFICANT ASSOCIATIONS IN THE CHROMOSOME 15 CHRNA3/5 REGION

Reference	Phenotype	Study Population	SNP with Most Significant Association	Evidence for Association
Pillai (2009) (11)	COPD	Norway: 823 cases and 810 controls	rs8034191	$P = 0.0001$
Pillai (2009) (11)	COPD	ICGN: 1,891 siblings in 606 families	rs1051730	$P = 7 \times 10^{-6}$
Pillai (2009) (11)	COPD	NETT-NAS: 389 cases and 472 controls	rs8034191	$P = 0.003$
Amos (2008) (12)	Lung cancer	Total of 3,878 cases and 4,831 controls	rs8034191	Combined $P = 3 \times 10^{-18}$
Hung (2008) (13)	Lung cancer	Total of 4,435 cases and 7,272 controls	rs8034191	Combined $P = 5 \times 10^{-20}$
Thorgeirsson (2008) (14)	Lung cancer	Total of 1,024 cases and 32,244 controls	rs1051730	Combined $P = 2 \times 10^{-8}$
Thorgeirsson (2008) (14)	Peripheral arterial disease	Total of 2,738 cases and 29,964 controls	rs1051730	Combined $P = 1 \times 10^{-7}$
Thorgeirsson (2008) (14)	Nicotine addiction (cigarettes per day as a categorical variable)	15,771 subjects	rs1051730	Combined $P = 6 \times 10^{-20}$
Berrettini (2008) (17)	Nicotine addiction (cigarettes per day)	5,634 Lausanne residents 1,847 subjects selected based on cholesterol and triglyceride levels	rs6495308	$P = 0.0006$ $P = 0.009$

Definition of abbreviations: COPD = chronic obstructive pulmonary disease; ICGN = International COPD Genetics Network; NETT-NAS = National Emphysema Treatment Trial–Normative Aging Study; SNP = single nucleotide polymorphism.

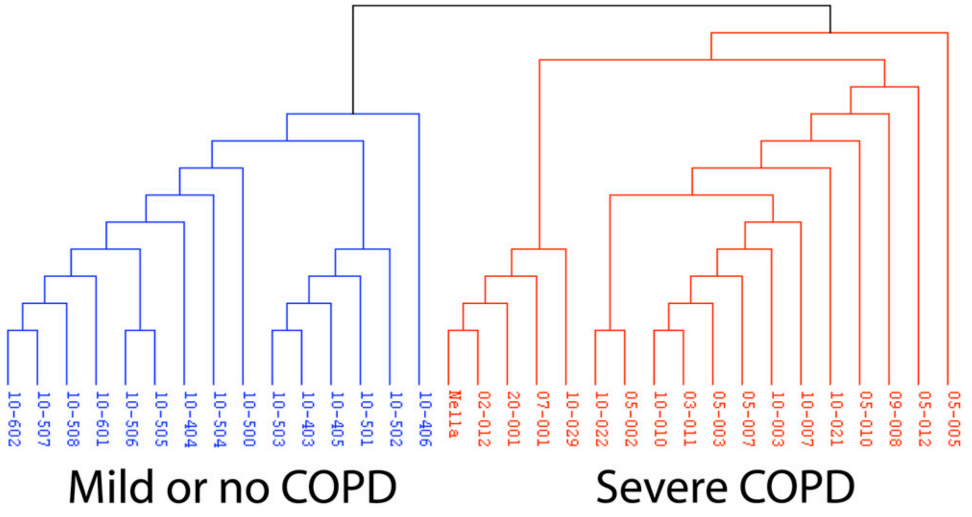


Figure 1. Hierarchical clustering of the lung tissue samples from the Spira and coworkers (18) study according to the expression of an 84 probeset biomarker derived from the study of Bhattacharya and colleagues (22). Dendogram of the samples from the study by Spira and coworkers is shown. The biomarker genes derived from the Bhattacharya study are able to accurately distinguish severe chronic obstructive pulmonary disease versus normal lung tissue in that independent cohort. Reproduced by permission from Reference 22.

types may provide novel insights into COPD pathophysiology.

GENE EXPRESSION STUDIES

Whole-genome gene expression studies of diseased lung and airway tissue provide an opportunity to gain an unbiased portrait of the COPD molecular landscape, offering insights into the processes that contribute to disease pathogenesis. While this rich molecular understanding of COPD could ultimately enable development of novel therapeutic approaches to target these pathogenic processes, we remain at the beginning of this journey. There are presently very few whole-genome expression studies of COPD lung tissue in humans (18–22). Some of the initial enthusiasm for using genome-wide gene expression in the setting of COPD has been diminished by the apparent lack of overlap, among these early studies, for genes differentially expressed in COPD lung tissue (21).

Two recent studies have shed some much-needed light on the potential relationships between existing gene expression studies of COPD lung tissue. They have attempted to overcome the inherent challenges in integrating results from relatively small studies that differ in inclusion/exclusion criteria, expression platform, and approaches for data analysis. Zeskind and coworkers (23) found that there was a high degree of overlap in the biological themes represented by the genes identified in the individual studies. While the differentially expressed genes are largely nonoverlapping, this study was able to use Gene Set Enrichment Analysis (GSEA) (24) to demonstrate that the gene-expression results from each paper share fundamental similarities. In a recently published study, Bhattacharya and colleagues (22) identified 220 genes whose expression in lung tissue was associated with COPD-related phenotypes in a cohort of 56 subjects with and without COPD. Importantly, they showed that a subset of these transcripts could serve as a biomarker distinguishing lung tissue from smokers with severe COPD versus mild or no COPD in a previously published gene-expression dataset (18) (Figure 1).

One exciting development has been the study of airway epithelial gene expression profiles in smokers with and without COPD. This work has been motivated by the concept that smoking creates a "field of molecular injury" in epithelial cells lining the entire respiratory tract, and that airway epithelial gene expression reflects host response to, and damage from, cigarette smoke (25–27). Heterogeneity in this expression pattern has been associated with tobacco-associated lung cancer (28, 29) and may provide insights into COPD-related processes. Pierrou and coworkers (30, 13) identified 200 oxidative stress-related genes that were differentially expressed in the bronchial airways between smokers with and without COPD. More recently, Tilley and colleagues (31) demonstrated that several Notch ligands, receptors, and downstream effector genes are down-regulated in the small airways (10th–12th generation) of smokers with COPD. Interestingly, Pierrou and coworkers found that a sizable fraction of the genes with COPD-specific patterns of differential expression in the airway were also differentially expressed in a previously published COPD lung tissue dataset (18). The ability to identify COPD-related processes in airway gene expression raises the possibility of developing airway biomarkers that could be used clinically to identify smokers at higher risk for developing disease as well as to serve as intermediate biomarkers of efficacy for novel and existing COPD therapies.

While there have been relatively few whole-genome expression studies of lung tissue and airways in COPD, some common themes are emerging across datasets to offer insights into disease pathogenesis. Larger studies on well-characterized cohorts of smokers with and without COPD using newer generations of microarrays and sequencing technology are needed to overcome many of the limitations of these early studies. Integration of these transcriptomic studies with whole-genome microRNA, methylation, and genotyping studies will ultimately yield a comprehensive molecular atlas of COPD.

***Conflict of Interest Statement*:** E.K.S. served as a consultant for GlaxoSmithKline (GSK) and AstraZeneca (AZ) $10,001–$50,000. He has received lecture fees from GSK $1,001–$5,000 and AZ $5,001–$10,000 and has received grant support from GSK, the NIH $100,001–more, and the American Lung Association $10,001–$50,000. A.S. served as a consultant for AllegroDx Inc. $10,001–$50,000 and also served on their Board or Advisory Board. He owns stocks or options of AllegroDx Inc. and has received grant support from the NIH $100,001–more. P.D.P. served on the Advisory Board for Talecris Biotherapeutics and received grant support from GSK, Merck, $100,001 or more, the NIH $50,001–$100,000, CIHR (Canada), and AllerGen NCE $100,001 or more.

References

1. Sandford AJ, Chagani T, Weir TD, Connett JE, Anthonisen NR, Paré PD. Susceptibility genes for rapid decline of lung function in the lung health study. *Am J Respir Crit Care Med* 2001;163:469–473.

2. Cheng SL, Yu CJ, Chen CJ, Yang PC. Genetic polymorphism of epoxide hydrolase and glutathione S-transferase in COPD. *Eur Respir J* 2004; 23:818–824.
3. van Diemen CC, Postma DS, Vonk JM, Bruinenberg M, Schouten JP, Boezen HM. A disintegrin and metalloprotease 33 polymorphisms and lung function decline in the general population. *Am J Respir Crit Care Med* 2005;172:329–333.
4. Aubier M, Guenegou A, Benessiano J, Leynaert B, Boczkowski J, Neukirch F. Association of lung function decline with the microsatellite polymorphism in the heme oxygenase-1 gene promoter, in a general population sample. Results from the Longitudinal European Community Respiratory Health Survey (ECRHS-France). *Bull Acad Natl Méd* 2006;190:877–890.
5. Imboden M, Downs SH, Senn O, Matyas G, Brändli O, Russi EW, Schindler C, Ackermann-Liebrich U, Berger W, Probst-Hensch NM. SAPALDIA Team. Glutathione S-transferase genotypes modify lung function decline in the general population: SAPALDIA cohort study. *Respir Res* 2007;8:2.
6. Seo T, Pahwa P, McDuffie HH, Yurube K, Egoshi M, Umemoto Y, Ghosh S, Fukushima Y, Nakagawa K. Association between cytochrome P450 3A5 polymorphism and the lung function in Saskatchewan grain workers. *Pharmacogenet Genomics* 2008;18:487–493.
7. Siedlinski M, Postma DS, van Diemen CC, Blokstra A, Smit HA, Boezen HM. Lung function loss, smoking, vitamin C intake, and polymorphisms of the glutamate-cysteine ligase genes. *Am J Respir Crit Care Med* 2008;178:13–19.
8. Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* 2005;6:95–108.
9. Eberle MA, Ng PC, Kuhn K, Zhou L, Peiffer DA, Galver L, Viaud-Martinez KA, Lawley CT, Gunderson KL, Shen R, *et al.* Power to detect risk alleles using genome-wide tag SNP panels. *PLoS Genet* 2007;3:1827–1837.
10. Manolio TA, Brooks LD, Collins FSA. HapMap harvest of insights into the genetics of common disease. *J Clin Invest* 2008;118:1590–1605.
11. Pillai SG, Ge D, Zhu G, Kong X, Shianna KV, Need AC, Feng S, Hersh CP, Bakke P, Gulsvik A,*et al.* A genome-wide association study in chronic obstructive pulmonary disease (COPD): identification of two major susceptibility loci. *PLoS Genet* 2009;5(3):e1000421.
12. Amos CI, Wu X, Broderick P, Gorlov IP, Gu J, Eisen T, Dong Q, Zhang Q, Gu X, Vijayakrishnan J, *et al.* Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. *Nat Genet* 2008;40:616–622.
13. Hung RJ, McKay JD, Gaborieau V, Boffetta P, Hashibe M, Zaridze D, Mukeria A, Szeszenia-Dabrowska N, Lissowska J, Rudnai P, *et al.* A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. *Nature* 2008;452:633–637.
14. Thorgeirsson TE, Geller F, Sulem P, Rafnar T, Wiste A, Magnusson KP, Manolescu A, Thorleifsson G, Stefansson H, Ingason A, *et al.* A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. *Nature* 2008;452:638–642.
15. Berrettini W, Yuan X, Tozzi F, Song K, Francks C, Chilcoat H, Waterworth D, Muglia P, Mooser V. Alpha-5/alpha-3 nicotinic receptor subunit alleles increase risk for heavy smoking. *Mol Psychiatry* 2008;13:368–373.
16. Young RP, Hopkins RJ, Hay BA, Epton MJ, Black PN, Gamble GD. Lung cancer gene associated with COPD: triple whammy or possible confounding effect? *Eur Respir J* 2008;32:1158–1164.
17. Wilk JB, Chen T-H, Gottlieb DJ, Walter RE, Nagle MW, Brandler BJ, Myers RH, Borecki IB, Silverman EK, Weiss ST, *et al.* A genome-wide association study of pulmonary function measurements in the Framingham Heart Study. *PLoS Genet* 2009; (In Press).
18. Spira A, Beane J, Pinto-Plata V, Kadar A, Liu G, Shah V, Celli B, Brody JS. Gene expression profiling of human lung tissue from smokers with severe emphysema. *Am J Respir Cell Mol Biol* 2004;31:601–610.
19. Golpon HA, Coldren CD, Zamora MR, Cosgrove GP, Moore MD, Tuder RM, Geraci MW, Voelkel NF. Emphysema lung tissue gene expression profiling. *Am J Respir Cell Mol Biol* 2004;31:595–600.
20. Ning W, Li CJ, Kaminski N, Feghali-Bostwick CA, Alber SM, Di YP, Otterbein SL, Song R, Hayashi S, Zhou Z, *et al.* Comprehensive gene expression profiles reveal pathways related to the pathogenesis of chronic obstructive pulmonary disease. *Proc Natl Acad Sci USA* 2004; 101:14895–14900.
21. Wang IM, Stepaniants S, Boie Y, Mortimer JR, Kennedy B, Elliott M, Hayashi S, Loy L, Coulter S, Cervino S, *et al.* Gene expression profiling in patients with chronic obstructive pulmonary disease and lung cancer. *Am J Respir Crit Care Med* 2008;177:402–411.
22. Bhattacharya S, Srisuma S, Demeo DL, Shapiro SD, Bueno R, Silverman EK, Reilly JJ, Mariani TJ. Molecular biomarkers for quantitative and discrete COPD phenotypes. *Am J Respir Cell Mol Biol* 2009;40: 359–367.
23. Zeskind JE, Lenburg ME, Spira A. Translating the COPD transcriptome: insights into pathogenesis and tools for clinical management. *Proc Am Thorac Soc* 2008;5:834–841. (Review).
24. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 2005;102:15545–15550.
25. Spira A, Beane J, Shah V, Liu G, Schembri F, Yang X, Palma J, Brody JS. Effects of cigarette smoke on the human airway epithelial cell transcriptome. *Proc Natl Acad Sci USA* 2004;101:10143–10148.
26. Chari R, Lonergan KM, Ng RT, MacAulay C, Lam WL, Lam S. Effect of active smoking on the human bronchial epithelium transcriptome. *BMC Genomics* 2007;8:297.
27. Hackett NR, Heguy A, Harvey BG, O'Connor TP, Luettich K, Flieder DB, Kaplan R, Crystal RG. Variability of antioxidant-related gene expression in the airway epithelium of cigarette smokers. *Am J Respir Cell Mol Biol* 2003;29:331–343.
28. Spira A, Beane JE, Shah V, Steiling K, Liu G, Schembri F, Gilman S, Dumas YM, Calner P, Sebastiani P, *et al.* Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer. *Nat Med* 2007;13:361–366.
29. Beane J, Sebastiani P, Whitfield TH, Steiling K, Dumas YM, Lenburg ME, Spira A. A prediction model for lung cancer diagnosis that integrates genomic and clinical features. *Cancer Prev Res (Phila Pa)* 2008;1:56–64.
30. Pierrou S, Broberg P, O'Donnell RA, Pawlowski K, Virtala R, Lindqvist E, Richter A, Wilson SJ, Angco G, Moller S, *et al.* Expression of genes involved in oxidative stress responses in airway epithelial cells of smokers with chronic obstructive pulmonary disease. *Am J Respir Crit Care Med* 2007;175:577–586.
31. Tilley AE, Harvey BG, Heguy A, Hackett NR, Wang R, O'Connor TP, Crystal RG. Down-regulation of the notch pathway in human airway epithelium in association with smoking and chronic obstructive pulmonary disease. *Am J Respir Crit Care Med* 2009;179:457–466.

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/262267846

Human genetics of tuberculosis: A long and winding road

Article *in* Philosophical Transactions B · June 2014

DOI: 10.1098/rstb.2013.0428 · Source: PubMed

CITATIONS
184

READS
658

5 authors, including:

Jamila EL Baghdadi
Military Hospital Mohammed V, Rabat
71 PUBLICATIONS **5,068** CITATIONS
SEE PROFILE

Ahmed Aziz Bousfiha
University of Hassan II Casablanca
140 PUBLICATIONS **2,716** CITATIONS
SEE PROFILE

Jean-Laurent Casanova
Rockefeller University
1,362 PUBLICATIONS **113,542** CITATIONS
SEE PROFILE

Erwin Schurr
McGill University
147 PUBLICATIONS **9,108** CITATIONS
SEE PROFILE

All content following this page was uploaded by Erwin Schurr on 22 July 2014.

The user has requested enhancement of the downloaded file.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

rstb.royalsocietypublishing.org

Review

Cite this article: Abel L, El-Baghdadi J, Bousfiha AA, Casanova J-L, Schurr E. 2014 Human genetics of tuberculosis: a long and winding road. *Phil. Trans. R. Soc. B* **369**: 20130428.
http://dx.doi.org/10.1098/rstb.2013.0428

One contribution of 12 to a Theme Issue 'After 2015: infectious diseases in a new era of health and development'.

Subject Areas:
genetics, immunology, health and disease and epidemiology

Keywords:
primary tuberculosis, pulmonary tuberculosis, latent tuberculosis infection, Mendelian predisposition, complex genetic predisposition, genetic variant

Authors for correspondence:
Laurent Abel
e-mail: laurent.abel@inserm.fr
Erwin Schurr
e-mail: erwin.schurr@mcgill.ca

Human genetics of tuberculosis: a long and winding road

Laurent Abel[1,2,3], Jamila El-Baghdadi[4], Ahmed Aziz Bousfiha[5], Jean-Laurent Casanova[1,2,3,6] and Erwin Schurr[7]

[1]Laboratory of Human Genetics of Infectious Diseases, Necker Branch, Institut National de la Santé et de la Recherche Médicale U1163, 75015 Paris, France
[2]Paris Descartes University, Sorbonne Paris Cité, Imagine Institute, 75015 Paris, France
[3]St Giles Laboratory of Human Genetics of Infectious Diseases, Rockefeller Branch, The Rockefeller University, New York, NY 10065, USA
[4]Genetics Unit, Military Hospital Mohamed V, Hay Riad, 10100 Rabat, Morocco
[5]Clinical Immunology Unit, Department of Pediatric Infectious Diseases, Hospital-University Center Ibn Rochd, King Hassan II University, Casablanca, Morocco
[6]Howard Hughes Medical Institute, The Rockefeller University, New York, NY 10065, USA
[7]McGill International TB Centre, The Research Institute of the McGill University Health Centre, Montreal, Quebec, Canada H3G 1A4

Only a small fraction of individuals exposed to *Mycobacterium tuberculosis* develop clinical tuberculosis (TB). Over the past century, epidemiological studies have shown that human genetic factors contribute significantly to this interindividual variability, and molecular progress has been made over the past decade for at least two of the three key TB-related phenotypes: (i) a major locus controlling resistance to infection with *M. tuberculosis* has been identified, and (ii) proof of principle that severe TB of childhood can result from single-gene inborn errors of interferon-γ immunity has been provided; genetic association studies with pulmonary TB in adulthood have met with more limited success. Future genetic studies of these three phenotypes could consider subgroups of subjects defined on the basis of individual (e.g. age at TB onset) or environmental (e.g. pathogen strain) factors. Progress may also be facilitated by further methodological advances in human genetics. Identification of the human genetic variants controlling the various stages and forms of TB is critical for understanding TB pathogenesis. These findings should have major implications for TB control, in the definition of improved prevention strategies, the optimization of vaccines and clinical trials and the development of novel treatments aiming to restore deficient immune responses.

1. Introduction

Tuberculosis (TB) remains a major public health problem, as *Mycobacterium tuberculosis* infects an estimated one-third of the world's population, resulting in approximately 8.6 million new cases of TB and approximately 1.3 million deaths in 2012 [1]. TB bacilli are transmitted by the inhalation of aerosolized droplets generated by the coughing of a patient with active TB. A substantial proportion of subjects do not become infected despite sustained high levels of exposure, as shown by negative tuberculin skin test (TST) and/or interferon (IFN)-γ release assays (IGRAs), and hence never develop disease (figure 1). About 5% of infected individuals develop clinical TB within 2 years of infection, either without latency or after a very short latent phase (figure 1) [2,3]. This 'primary' TB is particularly common in children, running an acute course and often associated with extrapulmonary disease owing to dissemination of the bacillus in the bloodstream [4,5]. However, most people infected with *M. tuberculosis* develop latent TB infection (LTBI). LTBI is characterized by positive TST and/or IGRA and an absence of overt clinical signs (figure 1) [2,6,7]. Most subjects with LTBI (approx. 90–95%) never develop clinical disease. The remaining 5–10% develop clinical TB later in life, typically owing to reactivation of the

© 2014 The Authors. Published by the Royal Society under the terms of the Creative Commons Attribution License http://creativecommons.org/licenses/by/3.0/, which permits unrestricted use, provided the original author and source are credited.

Royal Society Publishing

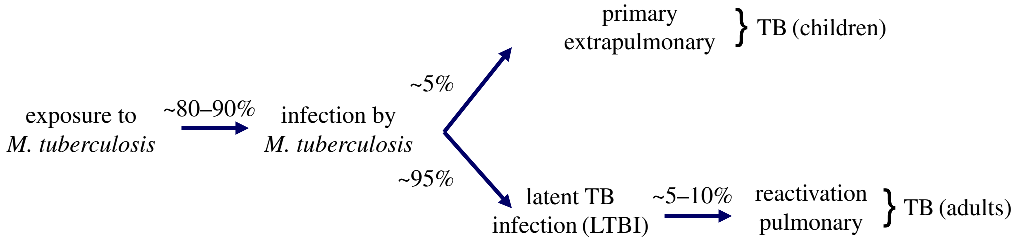


Figure 1. A schematic of the natural history of human infection by *M. tuberculosis*, and the subsequent development of clinical TB. Despite sustained high-level exposure, a substantial proportion of subjects (approx. 10–20%) do not become infected, and hence never develop disease. About 5% of infected individuals develop clinical TB within 2 years of infection; this 'primary' TB is particularly common in children, and could be associated with extrapulmonary disease. The remaining persons infected with *M. tuberculosis* develop latent TB infection (LTBI). Only a minority of subjects with LTBI (approx. 5–10%) develop clinical TB during their lifetime, typically owing to reactivation of the original infection. (Online version in colour.)

original infection. This 'reactivation' TB is predominantly a chronic pulmonary disease of adults, resulting in extensive lung damage and the efficient airborne transmission of bacteria. Hence, individuals susceptible to infection display three clinical presentations: (i) not entering LTBI (primary TB), (ii) remaining with LTBI (silent infection) or (iii) exiting from LTBI (reactivation TB).

Clinical and epidemiological surveys conducted since the 1910s have provided strong evidence that each step underlying infection or disease is controlled by host genetic factors [8–10]. Familial aggregation studies have provided the most convincing evidence [9,10]. In families with an index sputum-positive TB patient, spouses with family histories of TB were found to develop manifest TB more frequently than those without such histories [10]. Twin studies have also shown much higher concordance rates for monozygotic than for dizygotic pairs for clinical TB (combining primary and reactivation TB) [10,11]. In the seminal study conducted by Kallmann and Reisner in New York State, 308 twin pairs in which one of the pair was a confirmed index TB case were studied [11]. In this sample, the percentage of twin siblings of index cases developing manifest TB (defined on the basis of clinical, chest X-ray and sputum examinations) was 66.7% for monozygotic twins (52/78) and 23% (53/230) for dizygotic twins [11]. Interestingly, similar percentages were obtained when the sample was subdivided into groups on the basis of the known history of exposure of the twin siblings to a known active TB patient (table 1). It has long been known that the incidence of TB is particularly high in newly exposed populations [9], such as African populations and native Americans [12]. Similarly, the susceptibility to *M. tuberculosis* infection of an exposed individual, as measured by the TST, has been shown to be correlated with the region of ancestry of the individual concerned [13,14]. As for clinical TB disease, familial studies, including twin studies, have shown TB infection phenotypes (mostly TST result, considered as a quantitative trait) to be highly heritable (more than 50%, as detailed in §2).

Furthermore, a long series of experimental studies in various animal models, beginning in the 1930s, has also established the importance of host genetic background for determining the outcome of infection with *M. tuberculosis* (reviewed in [7,15,16]). In mice, a key antimycobacterial locus (the *Bcg* locus) has been shown to be allelic to *Nramp1* [17]. Knockout mice showed that CD4 T cells were required for immunity to *M. tuberculosis* and that the three most important antimycobacterial molecules required for *M. tuberculosis* destruction by phagocytes were tumour necrosis factor (TNFI-α), IFN-γ and nitric oxide synthase 2 [7,15,18,19]. Human studies have confirmed most of these findings through demonstrations of an increase in the risk of TB in AIDS patients [20,21], in patients on anti-TNF treatment [22], and in patients with genetic defects impairing IFN-γ immunity [8,23,24]. These studies have clearly demonstrated the critical importance of CD4-mediated immunity and of the interleukin (IL)-12/IFN-γ loop in baseline resistance to *M. tuberculosis*, but the reasons for which individuals otherwise displaying full immune competence develop TB remain largely unknown. In this context, the identification of genetic variants increasing the risk of TB constitutes a powerful approach to deciphering the mechanistic basis of TB pathogenesis. We provide here an overview of the principal human genes known to underlie the interindividual variability of susceptibility at each of the three main steps in the natural course of *M. tuberculosis* infection. We also discuss the contribution of genetic research to the development of new approaches to combat TB.

2. Genetic control of tuberculosis infection

There is no direct test for infection with *M. tuberculosis* and the phenotype of *M. tuberculosis* infection is inferred exclusively from quantitative measurements of antimycobacterial immunity. These assays cannot distinguish a possible anamnestic response to *M. tuberculosis* from persistent infection with the bacillus. The TST is the most widely used method [25]. The skin induration generated by the TST is caused by the accumulation of histiocytes and T cells around intradermal deposits of *M. tuberculosis* antigens. More recently, two *in vitro* blood assays, measuring either the secretion of IFN-γ by lymphocytes or the frequency of IFN-γ producing blood cells in response to *M. tuberculosis* antigens (IGRAs), have been developed [26]. TST and IGRAs assess different aspects of antimycobacterial immunity and are not fully concordant in predicting infection with *M. tuberculosis* [27]. Little or no reactivity in these tests in individuals exposed to *M. tuberculosis* is indicative of innate resistance to *M. tuberculosis* infection. In household studies, 30–50% of contacts with heavy short-term exposure do not become infected [28,29], revealing substantial heterogeneity in susceptibility to infection. We and others have focused on TST and IGRA results as quantitative traits and have shown heritability to be high for the results of both tests following exposure to *M. tuberculosis*. In Gambia, a study of healthy twins estimated the heritability of TST responsiveness and quantitative IGRA reactivity at 71% and 39%, respectively [30]. The heritability of quantitative TST reactivity in young healthy children exposed to an active TB case has been estimated at 92% in Chile [31]. In a South African familial

Table 1. Proportion of monozygotic and dizygotic twins of index cases with TB, as a function of the history of exposure of the co-twin to a known active TB patient, from the study by Kallmann & Reisner [11].

	monozygotic co-twins			dizygotic co-twins		
	number of TB cases	total number	% of TB cases	number of TB cases	total number	% of TB cases
history of exposure	36	52	69.2	46	175	26.3
without known exposure	16	26	61.5	7	55	12.7
total	52	78	66.7	53	230	23.0

sample from an area of hyperendemic TB disease, the heritability of quantitative IFN-γ release responses was estimated to be between 43 and 58%, depending on the nature of the stimulating antigen [32]. Likewise, the heritability of the frequency of antigen-specific IFN-γ^+CD4^+ and IFN-γ^+CD8^+ has been estimated at 53–74% [32]. A complex segregation analysis of TST reactivity in related household contacts of TB index cases in a Colombian population provided evidence for a major codominant gene accounting for approximately 65% of TST variability [33]. By contrast, a family study conducted in Uganda reported a lower estimate of heritability for IGRA responses (approx. 17%) [34]. However, as IGRA reactivity was adjusted for TST response in this study, this lower estimate of heritability may reflect genetic components common to the TST and IGRA responses. More recent data from Uganda carefully adjusted for shared environment yielded higher estimates for the heritability of IGRA response (approx. 30%) [35].

Despite the strong evidence in favour of an impact of genetic factors on the assays used for LTBI, only a small number of studies have aimed to identify the genetic variants underlying susceptibility to *M. tuberculosis* infection. TST reactivity was the phenotype studied in all these studies. Candidate gene association studies have focused on TST response as a binary trait, defined according to various thresholds (0, 5 or 10 mm). In a large study of 3622 TST-positive individuals and 244 TST-negative healthy controls in Ghana, an *IL10* promoter haplotype (−2849A/−1082A/−819C) was found to be significantly more frequent in TST-positive than in TST-negative subjects (15.3% versus 9.7%, odds ratio (OR) = 2.09 (1.2–3.5), $p = 0.012$) [36]. This haplotype was also associated with low levels of circulating IL-10, suggesting a role for IL-10 in the initial host response to *M. tuberculosis*. However, the similarity of age distribution between TST-negative and TST-positive subjects was not shown. Consistent with the findings for the sample from Ghana, the prevalence of TST negativity was found to be 1.5 times higher in individuals carrying the high-level IL-10-producing genotype GG at single nucleotide polymorphism (SNP) −1082A>G than in individuals carrying the AA and AG genotypes, in an indigenous population from Brazil [37]. Additional associations of cytokine genes with TST reactivity in the Brazilian sample have yet to be replicated.

In Uganda, a genome-wide (GW) linkage analysis reported results suggesting that persistent TST negativity (defined as a TST < 10 or 5 mm, according to age and HIV status) was linked to chromosomal regions 2q21–2q24 and 5p13–5q22 [38]. A study of TST reactivity in a sample of multiplex families from South Africa identified two major loci affecting TST-positivity *per se* (*TST1*) and the intensity of TST reactivity (*TST2*) [39]. *TST1* was identified by focusing on the phenotype of TST positivity versus TST negativity (i.e. TST = 0), and maps to chromosome 11p14 (lod score = 3.81, $p = 1.4 \times 10^{-5}$). A second phenotype studied was the size in millimetres of the skin induration in TST-tested subjects. The size of the induration, considered as a quantitative trait, was impacted by a locus on chromosome 5p15 which was termed *TST2* (lod score = 4.00, $p = 9 \times 10^{-6}$). The most parsimonious explanation for the role of these two loci is that *TST1* reflects innate resistance to infection with *M. tuberculosis* whereas *TST2* reflects T-cell-mediated antimycobacterial immune responses. Unexpectedly, it was subsequently discovered that a locus affecting the production of TNF by blood cells in response to bacillus Calmette–Guérin (BCG) and BCG plus IFN-γ, *TNF1*, is genetically indistinguishable from *TST1* [40]. This raises the exciting possibility that innate resistance to *M. tuberculosis* infection may involve a TNF-mediated effector mechanism. Such a possibility dovetails neatly with the function of TNF in macrophage activation during early stages of infection. No GW association study (GWAS) has yet been performed for *M. tuberculosis* infection.

3. Genetic control of severe primary tuberculosis

About 5% of infected individuals develop clinical TB within 2 years of infection (figure 1), either without latency or after a very short latent phase [3]. This 'primary' TB is particularly common in children, some of whom develop a haematogenous disseminated form (referred to here as 'severe primary TB') [2]. Severe primary TB was, by far, the most frequent form in children in areas of endemic disease before BCG vaccines and antimycobacterial antibiotics became available, resulting in high rates of mortality in children under the age of 2 years [9,23,41]. The risk of severe primary TB remains highly dependent on age at primary infection, decreasing from 10 to 20% for children under the age of 1 year to less than 0.5% for children over the age of 5 years [4,5]. These severe forms are mostly either miliary or affect the central nervous system (causing meningitis, in particular), and they remain life-threatening conditions [4,5]. BCG vaccination provides some protection against severe disseminated TB in childhood, but this protection is incomplete [42]. The development of antibiotics has greatly decreased childhood mortality due to TB, but more than 80 000 children still die from TB each year [43]. One of the fundamental unresolved questions in the field of childhood TB therefore concerns the nature of the predisposition to the development of severe clinical forms in only a minority of infected children. The findings obtained in the last decade have provided the first clue to the riddle, by showing that at least some cases of severe TB can be explained by single-gene inborn errors of immunity.

The first molecular evidence that childhood TB might reflect a Mendelian predisposition came from the observation of severe TB in children with classical primary immunodeficiencies (PIDs) [44]. In particular, a substantial number of children with chronic granulomatous disease (CGD) were diagnosed with severe TB in several countries [45–49]. In a recent survey investigating the occurrence of mycobacterial diseases in CGD patients, TB was found to be rather common in patients living in countries in which TB is endemic [50]. However, CGD is a rare disorder characterized by a high prevalence of multiple infectious diseases. Further progress towards an understanding of the genetics of severe TB came from the study of the syndrome of Mendelian susceptibility to mycobacterial diseases (MSMDs), which is defined by a selective vulnerability to weakly virulent mycobacterial species, such as BCG and environmental mycobacteria [51]. MSMD patients also often suffer from non-typhoidal, extra-intestinal salmonellosis. Since 1996, germline mutations in seven autosomal (*IFNGR1*, *IFNGR2*, *IL12B*, *IL12RB1*, *STAT1*, *IRF8*, *ISG15*) and two X-linked (*NEMO*, *CYBB*) genes have been discovered in MSMD patients [8,52–55]. High levels of locus and allelic heterogeneity have resulted in the definition of 17 different disorders, accounting for about half the known cases [56]. These defects are physiologically related, as they all result in an impairment of IFN-γ immunity. Several MSMD patients, particularly those with IFN-γR1 [57,58] and IL-12p40 [59] deficiencies, have been shown to suffer from infections due to both weakly virulent mycobacteria and *M. tuberculosis*, raising the question as to whether the TB observed in these patients could also be attributed to a monogenic predisposition.

The first answer to this question came when several siblings of MSMD patients carrying the same genetic defect as the index case were found to display severe TB as their sole infectious phenotype. This situation was initially observed in a child with partial IFN-γR1 deficiency, who was a sibling of an MSMD patient [60], and was subsequently observed in a male subject from a large multiplex X-linked kindred carrying a specific mutation of *CYBB* impairing the IFN-γ-dependent respiratory burst in macrophages [53]. However, the most common genetic defect identified in patients with severe TB to date is complete IL-12Rβ1 deficiency [61]. In one family, an IL-12Rβ1-deficient sister of a patient with MSMD developed abdominal TB [62]. Several children with severe TB and complete IL-12Rβ1 deficiency in the absence of a familial history of infections with weakly virulent mycobacteria have been identified [63,64]. In a more systematic search for *IL12RB1* mutations in 50 children with severe TB, two patients (4%) with complete IL-12Rβ1 deficiency were identified [24]. Overall, these results provided proof of principle for monogenic predisposition to severe TB, and raised the possibility that a substantial proportion of children with severe TB carry single-gene inborn errors of immunity. This proportion has been estimated at up to 45% by theoretical calculations [23], and can now be determined experimentally, by whole-exome and whole-genome sequencing. These findings have already paved the way for new treatments based on physiopathology. While until recently it was difficult to envision how point-of-care genotyping could be implemented in TB diagnosis, recent advances in hand-held PCR technology now suggest patient genotyping as a viable tool even in low- and middle-income countries [65,66]. The best example is provided by patients with IL-12Rβ1 deficiencies presenting TB owing to impaired IFN-γ production, for whom treatment with recombinant human IFN-γ, in addition to antimycobacterial drugs, has been shown to be effective [67].

4. Genetic control of pulmonary tuberculosis

Most people infected with *M. tuberculosis* present LTBI and do not develop primary TB (figure 1). Epidemiological studies have indicated that approximately 5–10% of individuals with LTBI go on to develop active TB during their lifetime, this risk decreasing with increasing time since infection [7,68]. Molecular epidemiology studies have shown that active TB due to reactivation of the original strain can occur decades after the initial infection [69]. The progression of infection within a subject from LTBI to pulmonary TB (PTB) reflects an impairment of host resistance to *M. tuberculosis*. This process may be triggered by acquired immunodeficiency, such as HIV infection or anti-TNF treatment. However, in subjects without overt immunodeficiency, the pathogenesis of reactivation remains unclear. As mentioned above, there is strong evidence that the development of PTB is influenced by host genetic factors. This genetic control is also likely to be different from that involved in primary TB [23,70]. Indeed, the study of genetic susceptibility to adulthood PTB has proved more difficult than that of susceptibility to severe childhood TB. In particular, no variants of genes from the IL12/IFN-γ circuit have been convincingly associated with PTB as yet. Current knowledge resulting from attempts to identify the genetic variants associated with PTB points to underlying heterogeneity, possibly due, at least in part, to the long-standing, multistep relationship between *M. tuberculosis* and its human host, through the natural history of PTB disease [71,72]. In particular, it is possible that specific subgroups of individuals with PTB have certain genetic risk factors, whereas other subgroups have other genetic risk factors. Such subgroups could potentially be defined on the basis of individual factors, such as clinical characteristics, or extrinsic factors, such as pathogen variability.

The vast majority of studies conducted to date to determine the molecular basis of PTB susceptibility have been association studies investigating the role of specific candidate genes. Most classical genetic association studies investigating PTB have focused on candidate genes, and a number of common risk variants have been reported in particular in immunity-related genes such as those encoding DC-SIGN, Toll-like receptors 1 and 2, vitamin D receptor, TNF, IL-1β or some HLA class II molecules [73]. However, there has been a lack of consistency between most of the reported results of independent studies [73,74]. One of the most convincing findings was the identification of associated polymorphisms of the *NRAMP1* gene (alias *SLC11A1*), the human orthologue of the murine *Nramp1* gene [17]. Following the initial association reported in a Gambian population [75], a meta-analysis of a large number of studies showed that several *NRAMP1* polymorphisms were significantly associated with PTB in African and Asian populations, but not in populations of European descent [76]. Two studies also provided evidence of a role for *NRAMP1* in early-onset TB. The first showed significant linkage to the *NRAMP1* gene in a large aboriginal Canadian family in which an outbreak of TB occurred [77]. In a follow-up study, common polymorphic alleles of *NRAMP1* were shown to be strong risk factors for TB (OR = 3.13 (1.54–6.25)) in children living in Texas, mostly of Hispanic and African origin, resulting in an allelic association acting in the opposite direction to that observed in adults [78]. These observations are consistent with the hypothesis that *NRAMP1* polymorphisms affect the speed of progression from infection to TB disease, accounting for the high frequency of some common alleles in patients with pediatric disease and the paucity of patients with the same alleles among TB cases with disease onset during adulthood. Overall, these studies provide strong support for a role of *NRAMP1* in TB, with an effect that is heterogeneous across populations, epidemiological settings and clinical phenotypes. They also highlight the importance of considering age at TB onset in these analyses. Generally, in the field of infectious diseases, stronger genetic effects are more pronounced in early-onset cases than in late-onset cases [70]. For example, in leprosy the association of genetic variants with disease has been shown repeatedly to be highly age-dependent [79,80].

The concept of stronger genetic effects associated with early-onset disease was further supported by the positional cloning of the first major locus conferring predisposition to PTB, which was found to be linked to chromosome region 8q [81]. Refined association mapping of the linked region identified variants of the *TOX* gene as strongly associated (OR = 3.09(1.99–4.78)) with the development of early-onset PTB (before 25 years of age) in populations from Morocco and Madagascar [82]. *TOX* encodes a nuclear factor involved in the development of T cells [83], particularly the CD4$^+$ T cells [84,85] critical for immunity to mycobacteria [20]. Conversely, GWASs on PTB, considered as a single phenotype, have met with limited success to date. A first GWAS performed on a large sample from Gambia and Ghana identified a single SNP with a weak effect (OR = 1.19 (1.12–1.26)) located in a ‘gene desert’ on chromosome 18 as a risk factor for PTB [86]. Further imputation of the original Ghanaian data identified a second locus on chromosome 11p13 as protective against TB [87]. The main protective SNP allele was well replicated in the Gambian sample but displayed only borderline associations in Indonesian patients and in a very large sample (more than 10 000 subjects) of subjects from Russia [87]. Another GWAS recently conducted in a South African population confirmed the protective effect of the chromosome 11p13 factor but identified no new risk loci of GW significance [88]. Finally, a GWAS in Asian populations identified an independent TB risk locus in chromosome region 20q12, only in patients defined as ‘young cases’ (OR = 1.73 (1.42–2.11)) [89]. Based on the observed age distribution of this latter study, young cases were defined as having an onset of TB before the age of 45 years, although more refined analyses with lower age thresholds showed that higher deviations of p-values from the null hypothesis were observed with younger age cut-offs despite the lower number of cases in each subset [89]. Overall, a striking feature of these GWASs is the lack of replication of the PTB susceptibility factors previously detected in candidate gene analyses [8,73,74]. Overall, the GWAS results suggest that common variants may have a limited impact on predisposition to adult PTB, at least when considered as a single phenotype, and point to underlying heterogeneity, possibly in phenotype definition.

5. Perspectives

TB was long considered to be purely infectious, but there is increasing evidence to suggest that this disease also reflects host genetic vulnerability. However, the precise nature of the genetic factors involved remains largely unknown. Several non-mutually exclusive explanations can account for the difficulties experienced in identifying the causal variants, especially in GWASs. Genetic heterogeneity may play a role, together with a complex mode of inheritance involving incomplete penetrance and modifier genes [90,91]. Likewise, the contribution of rare variants to TB pathogenesis is attracting attention for two main reasons: (i) conceptually, rare variants bridge the gap between Mendelian and complex inheritance, and may account for the major loci identified through linkage studies [56,70,92]; (ii) experimentally, these variants abound in the genome and can now be studied by whole-exome and whole-genome sequencing [93,94]. However, the phenotypic heterogeneity of TB-related traits causes more serious problems. For example, three measurements of immune reactivity are currently used to detect infection with *M. tuberculosis*. However, the results of these tests display limited concordance, and each assay captures a different aspect of the antimycobacterial response [27]. MSMD probably provides the best example of the critical interplay between phenotype and genetic control. The syndromic phenotype is MSMD, but the identification of the underlying genetic defects was greatly facilitated by establishing endophenotypes (e.g. specific patterns of cytokine production such as IL-12 and IFN-γ [95]), guiding subsequent genetic analysis. Endophenotypes are more closely related to gene function and are likely to be generally useful for the dissection of TB phenotypes. In PTB, the most commonly used phenotype-defining characteristic is the presence of *M. tuberculosis* in the sputum of patients, regardless of the other clinical, microbiological and demographic covariables. This approach completely ignores the dynamic nature of pulmonary TB, the likelihood of different stages of this process being under different genetic controls, as shown in a mouse model of BCG infection [96], and the possible impact of the *M. tuberculosis* strain on LTBI and clinical outcome [97–99]. The need for improvements in the definition of a more homogeneous and more refined TB phenotype for genetic studies is also demonstrated by the strong impact of age at onset of disease on our ability to detect genetic effects in clinical TB [78,82].

Traditional genetic approaches will also benefit from a better understanding of the molecular mechanisms of TB pathogenesis throughout the infectious cycle. For example, transcriptomic studies, including the search for loci involved in the expression of genes and/or microRNAs (expression quantitative trait loci, eQTL) in cell- or tissue-specific studies [100], based on RNA-seq in particular [101], are of major importance. A transcriptomic analysis of peripheral blood cells identified a neutrophil-driven IFN-α/β-inducible transcript signature in individuals with active PTB [102]. This blood signature was validated in independent studies [103–105], and was shown to be different from that of several infectious and pulmonary diseases [102,106], providing new insights into the most relevant pathways and candidate biomarkers for investigation in TB [107]. Some eQTLs associated with variation in gene expression levels in dendritic cells infected with *M. tuberculosis* [108] have also provided interesting candidate loci that remain to be tested for association with TB infection phenotypes. It is also possible that somatic mutations and epigenetic effects have a substantial impact on clinical susceptibility to TB, particularly in patients with late-onset disease [70]. Epigenetic mechanisms play a critical role in the immune system [109], and their investigation, particularly through GW analyses of the methylome [110,111], should provide new insight into the genetic basis of infection and clinical TB [112]. These epigenetic factors may be influenced by environmental and/or other host factors, such as diet, vitamin status or ageing [113,114], leading to additional possible gene × environment interaction effects. One of the main challenges in the next few years will be the integrated analysis of all these different sources of genomic information [115,116]. Finally, studies of various animal models, including zebrafish [90,91], rodents [15,16,117] and non-human primates [117,118], will be essential to provide critical information complementary to that obtained in human studies [7].

In the principal countries in which TB is endemic, disease control is based largely on passive case identification and drug treatment. This approach has proved effective for decreasing case mortality and human suffering, but the impact of current TB control on global TB trends is less clear [119–121]. The emergence and spread of extremely drug resistant strains resistant to the antibiotics currently available are warning signs that additional approaches will be required to halt endemic TB [122]. The availability of an effective vaccine against post-primary TB would provide us with a useful tool for decreasing TB transmission. Unfortunately, recent vaccine trials have yielded disappointing results [123,124], possibly reflecting our incomplete understanding of intrinsic vulnerability to TB as (i) mice that are genetically more resistant to *M. tuberculosis* infection benefit more from BCG vaccination than their susceptible counterparts [125], (ii) antimycobacterial immunity in humans is highly heritable and, thus, has a strong genetic component [30–32] and (iii) TB patients are more likely to suffer subsequent episodes of TB than would be expected on the basis of population incidence, strongly suggesting that TB patients have a susceptibility that cannot be overcome by conventional vaccination and/or antibiotics [126]. The strong positive selection of T-cell epitopes by *M. tuberculosis* across clinical strains presents an additional hurdle for vaccine development [127]. Taken together, these experimental data and observations *in natura* suggest that any effective TB vaccine must take into account the genetic susceptibility of patients if it is to trigger a protective response. Likewise, new therapeutic approaches based on the complementation of a specific immunodeficiency identified by human genetics should also have a positive complementary impact on the classical treatment of TB. For example, young patients developing disseminated TB during primary infection owing to the impaired production of IFN-γ (such as those with *IL12RB1* mutations) would benefit from targeted treatment with recombinant IFN-γ [67]. Finally, the recent discovery of *TST1* as a locus with an impact on intrinsic resistance to infection with *M. tuberculosis* provides another example for the usefulness of human genetics in TB control [39]. Once identified, the molecular mechanisms underlying *TST1* might constitute attractive targets for the prevention of infection by drug-based or vaccine interventions. The prevention of infection is the gold standard for stopping the TB endemic and studies of host genetics are an important weapon in the war against TB.

Acknowledgements. We thank A. Alcaïs, S. Boisson-Dupuis, J. Bustamante and all members of the Necker branch and Rockefeller branch of the laboratory of Human Genetics of Infectious Diseases.

Funding statement. This work was supported by grants from the European Research Council (ERC-2010-AdG-268777), the French National Agency for Research (ANR, grant no. ANR-08-MIEN-014-02), EU-grant HOMITB (HEALTH-F3-2008-200732), the Canadian Institutes of Health Research (CIHR), the St. Giles Foundation, the Rockefeller University grant no. 8ULTR000043 from the National Center for Research Resources and the National Center for Advancing Sciences (NCATS) of the National Institutes of Health, the Rockefeller University, the National Institute of Allergy and Infectious Diseases grant no. U01AI088685.

References

1. WHO. 2013 *Global tuberculosis report*. Geneva, Switzerland: World Health Organization.
2. Stewart GR, Robertson BD, Young DB. 2003 Tuberculosis: a problem with persistence. *Nat. Rev. Microbiol.* **1**, 97–105. (doi:10.1038/nrmicro749)
3. Zumla A, Raviglione M, Hafner R, von Reyn CF. 2013 Tuberculosis. *N Engl. J. Med.* **368**, 745–755. (doi:10.1056/NEJMra1200894)
4. Cruz AT, Starke JR. 2007 Clinical manifestations of tuberculosis in children. *Paediatr. Respir. Rev.* **8**, 107–117. (doi:10.1016/j.prrv.2007.04.008)
5. Marais BJ, Gie RP, Schaaf HS, Beyers N, Donald PR, Starke JR. 2006 Childhood pulmonary tuberculosis: old wisdom and new challenges. *Am. J. Respir. Crit. Care Med.* **173**, 1078–1090. (doi:10.1164/rccm.200511-1809SO)
6. Ernst JD. 2012 The immunological life cycle of tuberculosis. *Nat. Rev. Immunol.* **12**, 581–591. (doi:10.1038/nri3259)
7. O'Garra A, Redford PS, McNab FW, Bloom CI, Wilkinson RJ, Berry MP. 2013 The immune response in tuberculosis. *Annu. Rev. Immunol.* **31**, 475–527. (doi:10.1146/annurev-immunol-032712-095939)
8. Casanova JL, Abel L. 2002 Genetic dissection of immunity to mycobacteria: the human model. *Annu. Rev. Immunol.* **20**, 581–620. (doi:10.1146/annurev.immunol.20.081501.125851)
9. Dubos RJ, Dubos J. 1952 *The white plague; tuberculosis, man and society*, 1st edn, p. 277. Boston, MA: Little Brown.
10. Puffer R. 1944 *Familial susceptibility to tuberculosis; its importance as a public health problem*, p. 106. Cambridge, MA: Harvard University Press.
11. Kallmann FJ, Reisner D. 1943 Twin studies on the significance of genetic factors in tuberculosis. *Am. Rev. Tuberc.* **47**, 549–574.
12. Stead WW. 1997 The origin and erratic global spread of tuberculosis. How the past explains the present and is the key to the future. *Clin. Chest Med.* **18**, 65–77. (doi:10.1016/S0272-5231(05)70356-7)
13. Stead WW. 1992 Genetics and resistance to tuberculosis. Could resistance be enhanced by genetic engineering? *Ann. Intern. Med.* **116**, 937–941. (doi:10.7326/0003-4819-116-11-937)
14. Stead WW, Senner JW, Reddick WT, Lofgren JP. 1990 Racial differences in susceptibility to infection by *Mycobacterium tuberculosis*. *N. Engl. J. Med.* **322**, 422–427. (doi:10.1056/NEJM199002153220702)
15. Di Pietrantonio T, Schurr E. 2005 Mouse models for the genetic study of tuberculosis susceptibility. *Brief Funct. Genomics Proteomics* **4**, 277–292. (doi:10.1093/bfgp/4.3.277)
16. Fortin A, Abel L, Casanova JL, Gros P. 2007 Host genetics of mycobacterial diseases in mice and men: forward genetic studies of BCG-osis and tuberculosis.

Annu. Rev. Genomics Hum. Genet. **8**, 163–192. (doi:10.1146/annurev.genom.8.080706.092315)

17. Vidal SM, Malo D, Vogan K, Skamene E, Gros P. 1993 Natural resistance to infection with intracellular parasites: isolation of a candidate for *Bcg*. *Cell* **73**, 469–485. (doi:10.1016/0092-8674(93) 90135-D)
18. Cooper AM. 2009 Cell-mediated immune responses in tuberculosis. *Annu. Rev. Immunol.* **27**, 393–422. (doi:10.1146/annurev.immunol.021908.132703)
19. North RJ, Jung YJ. 2004 Immunity to tuberculosis. *Annu. Rev. Immunol.* **22**, 599–623. (doi:10.1146/annurev.immunol.22.012703.104635)
20. Cooper AM. 2009 T cells in mycobacterial infection and disease. *Curr. Opin. Immunol.* **21**, 378–384. (doi:10.1016/j.coi.2009.06.004)
21. Suthar AB *et al.* 2012 Antiretroviral therapy for prevention of tuberculosis in adults with HIV: a systematic review and meta-analysis. *PLoS Med.* **9**, e1001270. (doi:10.1371/journal.pmed.1001270)
22. Keane J, Gershon S, Wise RP, Mirabile-Levens E, Kasznica J, Schwieterman WD, Siegel JN, Braun MM. 2001 Tuberculosis associated with infliximab, a tumor necrosis factor alpha-neutralizing agent. *N. Engl. J. Med.* **345**, 1098–1104. (doi:10.1056/NEJMoa011110)
23. Alcais A, Fieschi C, Abel L, Casanova JL. 2005 Tuberculosis in children and adults: two distinct genetic diseases. *J. Exp. Med.* **202**, 1617–1621. (doi:10.1084/jem.20052302)
24. Boisson-Dupuis S *et al.* 2011 IL-12Rbeta1 deficiency in two of fifty children with severe tuberculosis from Iran, Morocco, and Turkey. *PLoS ONE* **6**, e18524. (doi:10.1371/journal.pone.0018524)
25. Reichman LB. 1979 Tuberculin skin testing. The state of the art. *Chest* **76**, 764–770. (doi:10.1378/chest.76.6.764)
26. Richeldi L. 2006 An update on the diagnosis of tuberculosis infection. *Am. J. Respir. Crit. Care Med.* **174**, 736–742. (doi:10.1164/rccm.200509-1516PP)
27. Gallant CJ *et al.* 2010 Tuberculin skin test and *in vitro* assays provide complementary measures of antimycobacterial immunity in children and adolescents. *Chest* **137**, 1071–1077. (doi:10.1378/chest.09-1852)
28. Jereb J, Etkind SC, Joglar OT, Moore M, Taylor Z. 2003 Tuberculosis contact investigations: outcomes in selected areas of the United States, 1999. *Int. J. Tuberc. Lung Dis.* **7**, S384–S390.
29. Marks SM, Taylor Z, Qualls NL, Shrestha-Kuwahara RJ, Wilce MA, Nguyen CH. 2000 Outcomes of contact investigations of infectious tuberculosis patients. *Am. J. Respir. Crit. Care Med.* **162**, 2033–2038. (doi:10.1164/ajrccm.162.6.2004022)
30. Jepson A, Fowler A, Banya W, Singh M, Bennett S, Whittle H, Hill AV. 2001 Genetic regulation of acquired immune responses to antigens of *Mycobacterium tuberculosis*: a study of twins in West Africa. *Infect. Immun.* **69**, 3989–3994. (doi:10.1128/IAI.69.6.3989-3994.2001)
31. Sepulveda RL, Heiba IM, King A, Gonzalez B, Elston RC, Sorensen RU. 1994 Evaluation of tuberculin reactivity in BCG-immunized siblings. *Am. J. Respir. Crit. Care Med.* **149**, 620–624. (doi:10.1164/ajrccm.149.3.8118628)
32. Cobat A *et al.* 2010 High heritability of antimycobacterial immunity in an area of hyperendemicity for tuberculosis disease. *J. Infect Dis.* **201**, 15–19. (doi:10.1086/648611)
33. Cobat A, Barrera LF, Henao H, Arbelaez P, Abel L, Garcia LF, Schurr E, Alcais A. 2012 Tuberculin skin test reactivity is dependent on host genetic background in Colombian tuberculosis household contacts. *Clin. Infect. Dis.* **54**, 968–971. (doi:10.1093/cid/cir972)
34. Stein CM, Guwatudde D, Nakakeeto M, Peters P, Elston RC, Tiwari HK, Mugerwa R, Whalen CC. 2003 Heritability analysis of cytokines as intermediate phenotypes of tuberculosis. *J. Infect Dis.* **187**, 1679–1685. (doi:10.1086/375249)
35. Tao L, Zalwango S, Chervenak K, Thiel B, Malone LL, Qiu F, Mayanja-Kizza H, Boom WH, Stein CM. 2013 Genetic and shared environmental influences on interferon-gamma production in response to *Mycobacterium tuberculosis* antigens in a Ugandan population. *Am. J. Trop. Med. Hyg.* **89**, 169–173. (doi:10.4269/ajtmh.12-0670)
36. Thye T *et al.* 2009 IL10 haplotype associated with tuberculin skin test response but not with pulmonary TB. *PLoS ONE* **4**, e5420. (doi:10.1371/journal.pone.0005420)
37. Zembrzuski VM, Basta PC, Callegari-Jacques SM, Santos RV, Coimbra CE, Salzano FM, Hutz MH. 2010 Cytokine genes are associated with tuberculin skin test response in a native Brazilian population. *Tuberculosis* (*Edinb*) **90**, 44–49. (doi:10.1016/j.tube.2009.11.002)
38. Stein CM *et al.* 2008 Genome scan of *M. tuberculosis* infection and disease in Ugandans. *PLoS ONE* **3**, e4094. (doi:10.1371/journal.pone.0004094)
39. Cobat A *et al.* 2009 Two loci control tuberculin skin test reactivity in an area hyperendemic for tuberculosis. *J. Exp. Med.* **206**, 2583–2591. (doi:10.1084/jem.20090892)
40. Cobat A *et al.* 2013 Identification of a major locus, *TNF1*, that controls BCG-triggered tumor necrosis factor production by leukocytes in an area hyperendemic for tuberculosis. *Clin. Infect. Dis.* **57**, 963–970. (doi:10.1093/cid/cit438)
41. Ranke KE. 1910 Diagnose und epidemiologie des lungentuberculose des kindes. *Arch. Kinderheilkd* **54**, 279–306.
42. Maartens G, Wilkinson RJ. 2007 Tuberculosis. *Lancet* **370**, 2030–2043. (doi:10.1016/S0140-6736(07)61262-8)
43. WHO. 2011 *Global tuberculosis control*. Geneva, Switzerland: WHO/HTM/TB/2011.
44. Reichenbach J, Rosenzweig S, Doffinger R, Dupuis S, Holland SM, Casanova JL. 2001 Mycobacterial diseases in primary immunodeficiencies. *Curr. Opin. Allergy Clin. Immunol.* **1**, 503–511. (doi:10.1097/00130832-200112000-00003)
45. Bustamante J *et al.* 2007 BCG-osis and tuberculosis in a child with chronic granulomatous disease. *J. Allergy Clin. Immunol.* **120**, 32–38. (doi:10.1016/j.jaci.2007.04.034)
46. Lau YL, Chan GC, Ha SY, Hui YF, Yuen KY. 1998 The role of phagocytic respiratory burst in host defense against *Mycobacterium tuberculosis*. *Clin. Infect. Dis.* **26**, 226–227. (doi:10.1086/517036)
47. Lee PP *et al.* 2008 Susceptibility to mycobacterial infections in children with X-linked chronic granulomatous disease: a review of 17 patients living in a region endemic for tuberculosis. *Pediatr. Infect Dis. J* **27**, 224–230. (doi:10.1097/INF.0b013e31815b494c)
48. Martire B *et al.* 2008 Clinical features, long-term follow-up and outcome of a large cohort of patients with Chronic Granulomatous Disease: an Italian multicenter study. *Clin. Immunol.* **126**, 155–164. (doi:10.1016/j.clim.2007.09.008)
49. Movahedi M *et al.* 2004 Chronic granulomatous disease: a clinical survey of 41 patients from the Iranian primary immunodeficiency registry. *Int. Arch. Allergy Immunol.* **134**, 253–259. (doi:10.1159/000078774)
50. Conti *et al.* Submitted. Mycobacterial diseases in 71 patients with chronic granulomatous disease.
51. Al-Muhsen S, Casanova JL. 2008 The genetic heterogeneity of Mendelian susceptibility to mycobacterial diseases. *J. Allergy Clin. Immunol.* **122**, 1043–1051; quiz 1052–1043. (doi:10.1016/j.jaci.2008.10.037)
52. Bogunovic D *et al.* 2012 Mycobacterial disease and impaired IFN-gamma immunity in humans with inherited ISG15 deficiency. *Science* **337**, 1684–1688. (doi:10.1126/science.1224026)
53. Bustamante J *et al.* 2011 Germline *CYBB* mutations that selectively affect macrophages in kindreds with X-linked predisposition to tuberculous mycobacterial disease. *Nat. Immunol.* **12**, 213–221. (doi:10.1038/ni.1992)
54. Filipe-Santos O *et al.* 2006 Inborn errors of IL-12/23- and IFN-gamma-mediated immunity: molecular, cellular, and clinical features. *Semin. Immunol.* **18**, 347–361. (doi:10.1016/j.smim.2006.07.010)
55. Hambleton S *et al.* 2011 *IRF8* mutations and human dendritic-cell immunodeficiency. *N. Engl. J. Med.* **365**, 127–138. (doi:10.1056/NEJMoa1100066)
56. Casanova JL, Abel L. 2013 The genetic theory of infectious diseases: a brief history and selected illustrations. *Annu. Rev. Genomics Hum. Genet.* **14**, 215–243. (doi:10.1146/annurev-genom-091212-153448)
57. Dorman SE *et al.* 2004 Clinical features of dominant and recessive interferon gamma receptor 1 deficiencies. *Lancet* **364**, 2113–2121. (doi:10.1016/S0140-6736(04)17552-1)
58. Sasaki Y, Nomura A, Kusuhara K, Takada H, Ahmed S, Obinata K, Hamada K, Okimoto Y, Hara T. 2002 Genetic basis of patients with bacille Calmette-Guerin osteomyelitis in Japan: identification of dominant partial interferon-gamma receptor 1 deficiency as a predominant type. *J. Infect Dis.* **185**, 706–709. (doi:10.1086/339011)
59. Picard C *et al.* 2002 Inherited interleukin-12 deficiency: *IL12B* genotype and clinical phenotype of 13 patients from six kindreds. *Am. J. Hum. Genet.* **70**, 336–348. (doi:10.1086/338625)
60. Jouanguy E *et al.* 1997 Partial interferon-gamma receptor 1 deficiency in a child with tuberculoid

bacillus Calmette-Guerin infection and a sibling with clinical tuberculosis. *J. Clin. Invest.* **100**, 2658–2664. (doi:10.1172/JCI119810)

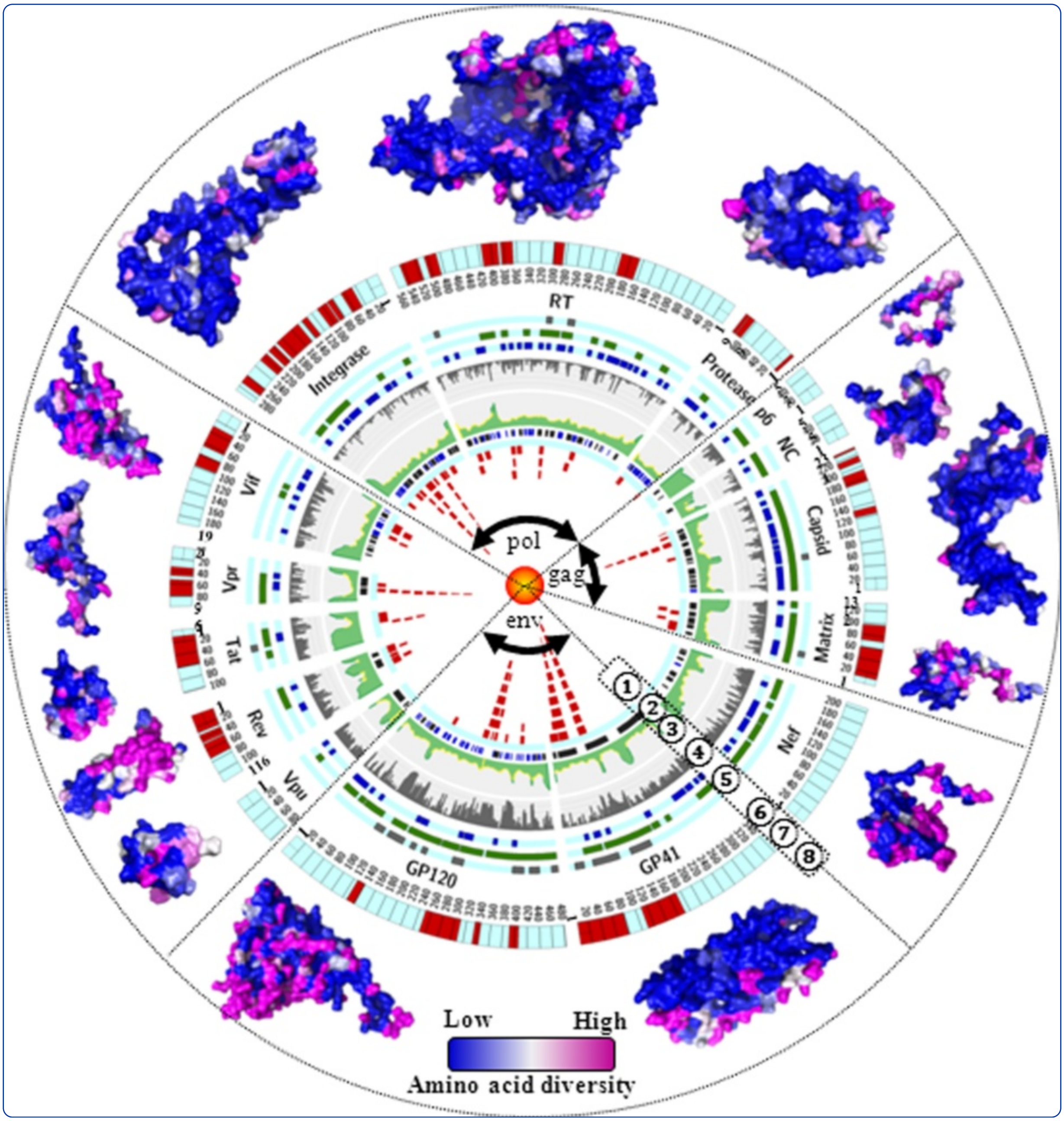
61. de Beaucoudrey L *et al*. 2010 Revisiting human IL-12Rbeta1 deficiency: a survey of 141 patients from 30 countries. *Medicine* (*Baltimore*) **89**, 381–402. (doi:10.1097/MD.0b013e3181fdd832)
62. Altare F *et al*. 2001 Interleukin-12 receptor beta1 deficiency in a patient with abdominal tuberculosis. *J. Infect Dis.* **184**, 231–236. (doi:10.1086/321999)
63. Caragol I *et al*. 2003 Clinical tuberculosis in 2 of 3 siblings with interleukin-12 receptor beta1 deficiency. *Clin. Infect. Dis.* **37**, 302–306. (doi:10.1086/375587)
64. Ozbek N, Fieschi C, Yilmaz BT, de Beaucoudrey L, Demirhan B, Feinberg J, Bikmaz YE, Casanova JL. 2005 Interleukin-12 receptor beta 1 chain deficiency in a child with disseminated tuberculosis. *Clin. Infect. Dis.* **40**, e55–e58. (doi:10.1086/427879)
65. Myers FB, Henrikson RH, Bone J, Lee LP. 2013 A handheld point-of-care genomic diagnostic system. *PLoS ONE* **8**, e70266. (doi:10.1371/journal.pone.0070266)
66. Roberts JD *et al*. 2012 Point-of-care genetic testing for personalisation of antiplatelet treatment (RAPID GENE): a prospective, randomised, proof-of-concept trial. *Lancet* **379**, 1705–1711. (doi:10.1016/S0140-6736(12)60161-5)
67. Alangari AA, Al-Zamil F, Al-Mazrou A, Al-Muhsen S, Boisson-Dupuis S, Awadallah S, Kambal A, Casanova JL. 2011 Treatment of disseminated mycobacterial infection with high-dose IFN-gamma in a patient with IL-12Rbeta1 deficiency. *Clin. Dev. Immunol.* **2011**, 691956. (doi:10.1155/2011/691956)
68. Vynnycky E, Fine PE. 2000 Lifetime risks, incubation period, and serial interval of tuberculosis. *Am. J. Epidemiol.* **152**, 247–263. (doi:10.1093/aje/152.3.247)
69. Lillebaek T, Dirksen A, Vynnycky E, Baess I, Thomsen VO, Andersen AB. 2003 Stability of DNA patterns and evidence of *Mycobacterium tuberculosis* reactivation occurring decades after the initial infection. *J. Infect Dis.* **188**, 1032–1039. (doi:10.1086/378240)
70. Alcais A, Quintana-Murci L, Thaler DS, Schurr E, Abel L, Casanova JL. 2010 Life-threatening infectious diseases of childhood: single-gene inborn errors of immunity? *Ann. N Y Acad. Sci.* **1214**, 18–33. (doi:10.1111/j.1749-6632.2010.05834.x)
71. Comas I *et al*. 2013 Out-of-Africa migration and Neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans. *Nat. Genet.* **45**, 1176–1182. (doi:10.1038/ng.2744)
72. Gagneux S. 2012 Host–pathogen coevolution in human tuberculosis. *Phil. Trans. R. Soc. B* **367**, 850–859. (doi:10.1098/rstb.2011.0316)
73. Azad AK, Sadee W, Schlesinger LS. 2012 Innate immune gene polymorphisms in tuberculosis. *Infect. Immun.* **80**, 3343–3359. (doi:10.1128/IAI.00443-12)
74. Moller M, de Wit E, Hoal EG. 2010 Past, present and future directions in human genetic susceptibility to tuberculosis. *FEMS Immunol. Med. Microbiol.* **58**, 3–26. (doi:10.1111/j.1574-695X.2009.00600.x)
75. Bellamy R, Ruwende C, Corrah T, McAdam KP, Whittle HC, Hill AV. 1998 Variations in the *NRAMP1* gene and susceptibility to tuberculosis in West Africans. *N Engl. J. Med.* **338**, 640–644. (doi:10.1056/NEJM199803053381002)
76. Li HT, Zhang TT, Zhou YQ, Huang QH, Huang J. 2006 *SLC11A1* (formerly *NRAMP1*) gene polymorphisms and tuberculosis susceptibility: a meta-analysis. *Int. J. Tuberc. Lung Dis.* **10**, 3–12.
77. Greenwood CM, Fujiwara TM, Boothroyd LJ, Miller MA, Frappier D, Fanning EA, Schurr E, Morgan K. 2000 Linkage of tuberculosis to chromosome 2q35 loci, including *NRAMP1*, in a large aboriginal Canadian family. *Am. J. Hum. Genet.* **67**, 405–416. (doi:10.1086/303012)
78. Malik S *et al*. 2005 Alleles of the *NRAMP1* gene are risk factors for pediatric tuberculosis disease. *Proc. Natl Acad. Sci. USA* **102**, 12 183–12 188. (doi:10.1073/pnas.0503368102)
79. Alcais A *et al*. 2007 Stepwise replication identifies a low-producing lymphotoxin-alpha allele as a major risk factor for early-onset leprosy. *Nat. Genet.* **39**, 517–522. (doi:10.1038/ng2000)
80. Alter A *et al*. 2011 Human leukocyte antigen class I region single-nucleotide polymorphisms are associated with leprosy susceptibility in Vietnam and India. *J. Infect Dis.* **203**, 1274–1281. (doi:10.1093/infdis/jir024)
81. Baghdadi JE *et al*. 2006 An autosomal dominant major gene confers predisposition to pulmonary tuberculosis in adults. *J. Exp. Med.* **203**, 1679–1684. (doi:10.1084/jem.20060269)
82. Grant AV *et al*. 2013 Age-dependent association between pulmonary tuberculosis and common *TOX* variants in the 8q12–13 linkage region. *Am. J. Hum. Genet.* **92**, 407–414. (doi:10.1016/j.ajhg.2013.01.013)
83. Aliahmad P, Seksenyan A, Kaye J. 2012 The many roles of TOX in the immune system. *Curr. Opin. Immunol.* **24**, 173–177. (doi:10.1016/j.coi.2011.12.001)
84. Aliahmad P, Kadavallore A, de la Torre B, Kappes D, Kaye J. 2011 TOX is required for development of the CD4 T cell lineage gene program. *J. Immunol.* **187**, 5931–5940. (doi:10.4049/jimmunol.1101474)
85. Aliahmad P, Kaye J. 2008 Development of all CD4 T lineages requires nuclear factor TOX. *J. Exp. Med.* **205**, 245–256. (doi:10.1084/jem.20071944)
86. Thye T *et al*. 2010 Genome-wide association analyses identifies a susceptibility locus for tuberculosis on chromosome 18q11.2. *Nat. Genet.* **42**, 739–741. (doi:10.1038/ng.639)
87. Thye T *et al*. 2012 Common variants at 11p13 are associated with susceptibility to tuberculosis. *Nat. Genet.* **44**, 257–259. (doi:10.1038/ng.1080)
88. Chimusa ER, Zaitlen N, Daya M, Moller M, van Helden PD, Mulder NJ, Price AL, Hoal EG. 2013 Genome-wide association study of ancestry-specific TB risk in the South African coloured population. *Hum. Mol. Genet*. **23**, 796–809. (doi:10.1093/hmg/ddt462)
89. Mahasirimongkol S *et al*. 2012 Genome-wide association studies of tuberculosis in Asians identify distinct at-risk locus for young tuberculosis. *J. Hum. Genet.* **57**, 363–367. (doi:10.1038/jhg.2012.35)
90. Behr M, Schurr E, Gros P. 2010 TB: screening for responses to a vile visitor. *Cell* **140**, 615–618. (doi:10.1016/j.cell.2010.02.030)
91. Tobin DM *et al*. 2010 The *lta4h* locus modulates susceptibility to mycobacterial infection in zebrafish and humans. *Cell* **140**, 717–730. (doi:10.1016/j.cell.2010.02.013)
92. Abel L, Casanova JL. 2000 Genetic predisposition to clinical tuberculosis: bridging the gap between simple and complex inheritance. *Am. J. Hum. Genet.* **67**, 274–277. (doi:10.1086/303033)
93. Altshuler DM *et al*. 2010 Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58. (doi:10.1038/nature09298)
94. Cirulli ET, Goldstein DB. 2010 Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat. Rev. Genet.* **11**, 415–425. (doi:10.1038/nrg2779)
95. Feinberg J *et al*. 2004 Bacillus Calmette Guerin triggers the IL-12/IFN-gamma axis by an IRAK-4- and NEMO-dependent, non-cognate interaction between monocytes, NK, and T lymphocytes. *Eur. J. Immunol.* **34**, 3276–3284. (doi:10.1002/eji.200425221)
96. Di Pietrantonio T, Hernandez C, Girard M, Verville A, Orlova M, Belley A, Behr MA, Loredo-Osti JC, Schurr E. 2010 Strain-specific differences in the genetic control of two closely related mycobacteria. *PLoS Pathog.* **6**, e1001169. (doi:10.1371/journal.ppat.1001169)
97. Herb F *et al*. 2008 *ALOX5* variants associated with susceptibility to human pulmonary tuberculosis. *Hum. Mol. Genet.* **17**, 1052–1060. (doi:10.1093/hmg/ddm378)
98. Intemann CD *et al*. 2009 Autophagy gene variant *IRGM*–261T contributes to protection from tuberculosis caused by *Mycobacterium tuberculosis* but not by *M. africanum* strains. *PLoS Pathog.* **5**, e1000577. (doi:10.1371/journal.ppat.1000577)
99. Nicol MP, Wilkinson RJ. 2008 The clinical consequences of strain diversity in *Mycobacterium tuberculosis*. *Trans. R. Soc. Trop. Med. Hyg.* **102**, 955–965. (doi:10.1016/j.trstmh.2008.03.025)
100. Gregersen PK. 2012 Cell type-specific eQTLs in the human immune system. *Nat. Genet.* **44**, 478–480. (doi:10.1038/ng.2258)
101. Pickrell JK *et al*. 2010 Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**, 768–772. (doi:10.1038/nature08872)
102. Berry MP *et al*. 2010 An interferon-inducible neutrophil-driven blood transcriptional signature in human tuberculosis. *Nature* **466**, 973–977. (doi:10.1038/nature09247)
103. Cliff JM *et al*. 2013 Distinct phases of blood gene expression pattern through tuberculosis treatment reflect modulation of the humoral immune response. *J. Infect Dis*. **207**, 18–29. (doi:10.1093/infdis/jis499)
104. Maertzdorf J, Repsilber D, Parida SK, Stanley K, Roberts T, Black G, Walzl G, Kaufmann SH. 2011 Human gene expression profiles of susceptibility

and resistance in tuberculosis. *Genes Immun.* **12**, 15–22. (doi:10.1038/gene.2010.51)

105. Ottenhoff TH *et al*. 2012 Genome-wide expression profiling identifies type 1 interferon response pathways in active tuberculosis. *PLoS ONE* **7**, e45839. (doi:10.1371/journal.pone.0045839)
106. Bloom CI *et al*. 2013 Transcriptional blood signatures distinguish pulmonary tuberculosis, pulmonary sarcoidosis, pneumonias and lung cancers. *PLoS ONE* **8**, e70630. (doi:10.1371/journal.pone.0070630)
107. Berry MP, Blankley S, Graham CM, Bloom CI, O'Garra A. 2013 Systems approaches to studying the immune response in tuberculosis. *Curr. Opin. Immunol.* **25**, 579–587. (doi:10.1016/j.coi.2013.08.003)
108. Barreiro LB, Tailleux L, Pai AA, Gicquel B, Marioni JC, Gilad Y. 2012 Deciphering the genetic architecture of variation in the immune response to *Mycobacterium tuberculosis* infection. *Proc. Natl Acad. Sci. USA* **109**, 1204–1209. (doi:10.1073/pnas.1115761109)
109. Suarez-Alvarez B, Rodriguez RM, Fraga MF, Lopez-Larrea C. 2012 DNA methylation: a promising landscape for immune system-related diseases. *Trends Genet.* **28**, 506–514. (doi:10.1016/j.tig.2012.06.005)
110. Krueger F, Kreck B, Franke A, Andrews SR. 2012 DNA methylome analysis using short bisulfite sequencing data. *Nat. Methods* **9**, 145–151. (doi:10.1038/nmeth.1828)
111. Laird PW. 2010 Principles and challenges of genomewide DNA methylation analysis. *Nat. Rev. Genet.* **11**, 191–203. (doi:10.1038/nrg2732)
112. Esterhuyse MM, Linhart HG, Kaufmann SH. 2012 Can the battle against tuberculosis gain from epigenetic research? *Trends Microbiol.* **20**, 220–226. (doi:10.1016/j.tim.2012.03.002)
113. Chiacchiera F, Piunti A, Pasini D. 2013 Epigenetic methylations and their connections with metabolism. *Cell Mol. Life Sci.* **70**, 1495–1508. (doi:10.1007/s00018-013-1293-5)
114. Kaelin Jr WG, McKnight SL. 2013 Influence of metabolism on epigenetics and disease. *Cell* **153**, 56–69. (doi:10.1016/j.cell.2013.03.004)
115. Hawkins RD, Hon GC, Ren B. 2010 Next-generation genomics: an integrative approach. *Nat. Rev. Genet.* **11**, 476–486. (doi:10.1038/nrg2795)
116. Knight JC. 2013 Genomic modulators of the immune response. *Trends Genet.* **29**, 74–83. (doi:10.1016/j.tig.2012.10.006)
117. O'Toole R. 2010 Experimental models used to study human tuberculosis. *Adv. Appl. Microbiol.* **71**, 75–89. (doi:10.1016/S0065-2164(10)71003-0)
118. Flynn JL. 2006 Lessons from experimental *Mycobacterium tuberculosis* infections. *Microbes Infect.* **8**, 1179–1188. (doi:10.1016/j.micinf.2005.10.033)
119. Dye C, Lonnroth K, Jaramillo E, Williams BG, Raviglione M. 2009 Trends in tuberculosis incidence and their determinants in 134 countries. *Bull World Health Organ.* **87**, 683–691. (doi:10.2471/BLT.08.058453)
120. Oxlade O, Schwartzman K, Behr MA, Benedetti A, Pai M, Heymann J, Menzies D. 2009 Global tuberculosis trends: a reflection of changes in tuberculosis control or in population health? *Int. J. Tuberc. Lung Dis.* **13**, 1238–1246.
121. Oxlade O, Schwartzman K, Benedetti A, Pai M, Heymann J, Menzies D. 2011 Developing a tuberculosis transmission model that accounts for changes in population health. *Med. Decis. Making* **31**, 53–68. (doi:10.1177/0272989X10369001)
122. Lienhardt C, Glaziou P, Uplekar M, Lonnroth K, Getahun H, Raviglione M. 2012 Global tuberculosis control: lessons learnt and future prospects. *Nat. Rev. Microbiol.* **10**, 407–416. (doi:10.1038/nrmicro2797)
123. Ota MO *et al*. 2011 Immunogenicity of the tuberculosis vaccine MVA85A is reduced by coadministration with EPI vaccines in a randomized controlled trial in Gambian infants. *Sci. Transl. Med.* **3**, 88ra56. (doi:10.1126/scitranslmed.3002461)
124. Tameris MD *et al*. 2013 Safety and efficacy of MVA85A, a new tuberculosis vaccine, in infants previously vaccinated with BCG: a randomised, placebo-controlled phase 2b trial. *Lancet* **381**, 1021–1028. (doi:10.1016/S0140-6736(13)60177-4)
125. Medina E, North RJ. 1999 Genetically susceptible mice remain proportionally more susceptible to tuberculosis after vaccination. *Immunology* **96**, 16–21. (doi:10.1046/j.1365-2567.1999.00663.x)
126. Verver S *et al*. 2005 Rate of reinfection tuberculosis after successful treatment is higher than rate of new tuberculosis. *Am. J. Respir. Crit. Care Med.* **171**, 1430–1435. (doi:10.1164/rccm.200409-1200OC)
127. Comas I, Chakravartti J, Small PM, Galagan J, Niemann S, Kremer K, Ernst JD, Gagneux S. 2010 Human T cell epitopes of *Mycobacterium tuberculosis* are evolutionarily hyperconserved. *Nat. Genet.* **42**, 498–503. (doi:10.1038/ng.590)

View publication stats

RETROVIROLOGY


RT
Integrase
Protease
p6
NC
Capsid
Matrix
Nef
GP41
GP120
Vpu
Rev
Tat
Vpr
Vif
19
116
pol
gag
env
①
②
③
④
⑤
⑥
⑦
⑧
Low
High
Amino acid diversity

An integrated map of HIV genome-wide variation from a population perspective

Li *et al.*

BioMed Central

Li *et al. Retrovirology* (2015) 12:18
DOI 10.1186/s12977-015-0148-6

Li *et al. Retrovirology* (2015) 12:18
DOI 10.1186/s12977-015-0148-6

RETROVIROLOGY

RESEARCH Open Access

An integrated map of HIV genome-wide variation from a population perspective

Guangdi Li[1,2*], Supinya Piampongsant[2], Nuno Rodrigues Faria[3], Arnout Voet[4], Andrea-Clemencia Pineda-Peña[2,5], Ricardo Khouri[2,6], Philippe Lemey[2], Anne-Mieke Vandamme[2,7] and Kristof Theys[2*]

Abstract

Background: The HIV pandemic is characterized by extensive genetic variability, which has challenged the development of HIV drugs and vaccines. Although HIV genomes have been classified into different types, groups, subtypes and recombinants, a comprehensive study that maps HIV genome-wide diversity at the population level is still lacking to date. This study aims to characterize HIV genomic diversity in large-scale sequence populations, and to identify driving factors that shape HIV genome diversity.

Results: A total of 2996 full-length genomic sequences from 1705 patients infected with 16 major HIV groups, subtypes and circulating recombinant forms (CRFs) were analyzed along with structural, immunological and peptide inhibitor information. Average nucleotide diversity of HIV genomes was almost 50% between HIV-1 and HIV-2 types, 37.5% between HIV-1 groups, 14.7% between HIV-1 subtypes, 8.2% within individual HIV-1 subtypes and less than 1% within single patients. Along the HIV genome, diversity patterns and compositions of nucleotides and amino acids were highly similar across different groups, subtypes and CRFs. Current HIV-derived peptide inhibitors were predominantly derived from conserved, solvent accessible and intrinsically ordered structures in the HIV-1 subtype B genome. We identified these conserved regions in Capsid, Nucleocapsid, Protease, Integrase, Reverse transcriptase, Vpr and the GP41 N terminus as potential drug targets. In the analysis of factors that impact HIV-1 genomic diversity, we focused on protein multimerization, immunological constraints and HIV-human protein interactions. We found that amino acid diversity in monomeric proteins was higher than in multimeric proteins, and diversified positions were preferably located within human CD4 T cell and antibody epitopes. Moreover, intrinsic disorder regions in HIV-1 proteins coincided with high levels of amino acid diversity, facilitating a large number of interactions between HIV-1 and human proteins.

Conclusions: This first large-scale analysis provided a detailed mapping of HIV genomic diversity and highlighted drug-target regions conserved across different groups, subtypes and CRFs. Our findings suggest that, in addition to the impact of protein multimerization and immune selective pressure on HIV-1 diversity, HIV-human protein interactions are facilitated by high variability within intrinsically disordered structures.

Keywords: HIV genome, Genomic diversity, Conservation, Peptide inhibitor, HIV-human protein interaction, HIV phylogenetic tree, HIV inter- and inter-clade genetic diversity, Selective pressure, Protein multimerization, Protein intrinsic disorder

---

* Correspondence: liguangdi.research@gmail.com; Kristof.Theys@rega.kuleuven.be
[1]Metabolic Syndrome Research Center, the Second Xiangya Hospital, Central South University, Changsha, Hunan, China
[2]Rega Institute for Medical Research, Department of Microbiology and Immunology, KU Leuven, Leuven, Belgium
Full list of author information is available at the end of the article

BioMed Central

© 2015 Li et al.; licensee BioMed Central. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

Background

As the causative agent of AIDS, the Human Immunodeficiency Virus (HIV) represents a worldwide threat to public health and the economy. The HIV pandemic is characterized by extensive genomic diversity caused by multiple factors including multiple zoonotic transmissions into human populations, high rates of viral evolution and recombination [1]. HIV has two major types, HIV-1 and HIV-2, which are further divided into groups, subtypes and recombinant forms. Globally, over 90% of HIV infections belong to HIV-1 group M viruses, which have been classified into 9 subtypes (A-D, F-H, J, K) and more than 50 circulating recombinant forms (CRFs) [1]. The high genetic diversity of the HIV genome has challenged the development of drugs and vaccines [2].

The HIV genome contains nine genes that encode fifteen viral proteins (Additional file 1: Figure S1). Three major genes, *gag*, *pol* and *env*, code for structural proteins (Matrix, Capsid, Nucleocapsid, p6), enzymes (Protease, Reverse transcriptase (RT), Integrase) and envelope proteins (GP120, GP41), respectively. The remaining genes code for regulatory (Tat, Rev) and accessory proteins (Vif, Vpr, Vpu/Vpx, Nef) [3]. These viral proteins can exhibit multiple functions and interact with various human proteins during the viral life cycle [4,5].

During the past three decades, many antiviral inhibitors have been designed to prevent HIV replication by targeting different viral proteins [6]. These anti-HIV peptides and small-molecule inhibitors either act by blocking active sites of viral enzymes or interrupting protein interactions [6]. For instance, the fusion inhibitor T20 (Enfuvirtide, Fuzeon), a peptide derived from the GP41 heptad repeat region, can efficiently inhibit viral entry by interrupting interactions between the GP41 helices [7]. For all existing drug classes, mutations in the HIV genome can cause drug resistance [8]. Therefore, inhibitors have been preferentially developed to target conserved regions of different viral proteins [9]. HIV genetic diversity also challenges the development of a global HIV vaccine [10]. While the vaccine trial STEP was unable to show preventive vaccination in subtype B infected cohorts [11], the Thai trial RV144 showed for the first time that prime-boost vaccination provided a modest efficacy in patients infected with CRF01_AE [12]. For vaccine and drug design, it remains important to investigate the genomic diversity of different HIV groups, subtypes and CRFs at a population level.

Despite a large body of knowledge on different aspects of HIV pathogenesis, a large-scale analysis that reveals the genome-wide diversity within and between different HIV groups, subtypes and CRFs is still lacking. Although previous HIV genomic studies have reported subtype distribution, genetic variability, disease progression, evolutionary rate, positive selective pressure and the origin of HIV [11-27], most studies reported their findings using either reference genomes or small cohorts of less than 100 patients or sequences in a single subtype. HIV-1 subtype B which dominates infections in developed countries is the most studied subtype, largely due to historical reasons [28]. For instance, the adaptive evolution during acute infection was evaluated only in 11 individuals infected with HIV-1 subtype B [14]. In light of using HIV consensus sequences as vaccine candidates, an analysis on the genetic difference between consensus sequences and circulating strains was limited to subtypes B and C using less than 100 sequences [2], while other subtypes also prevail worldwide [29].

The last three decades have seen an accumulation of HIV data including full-length genomic sequences, protein crystal structures, HIV-human protein interactions, human T-cell epitope information and antiretroviral peptide inhibitors derived from the HIV genome. By integrating distinct but complementary sources of large-scale HIV datasets, this study aims to characterize HIV genome-wide diversity and to determine multiple factors that shape HIV genomic diversity.

Results

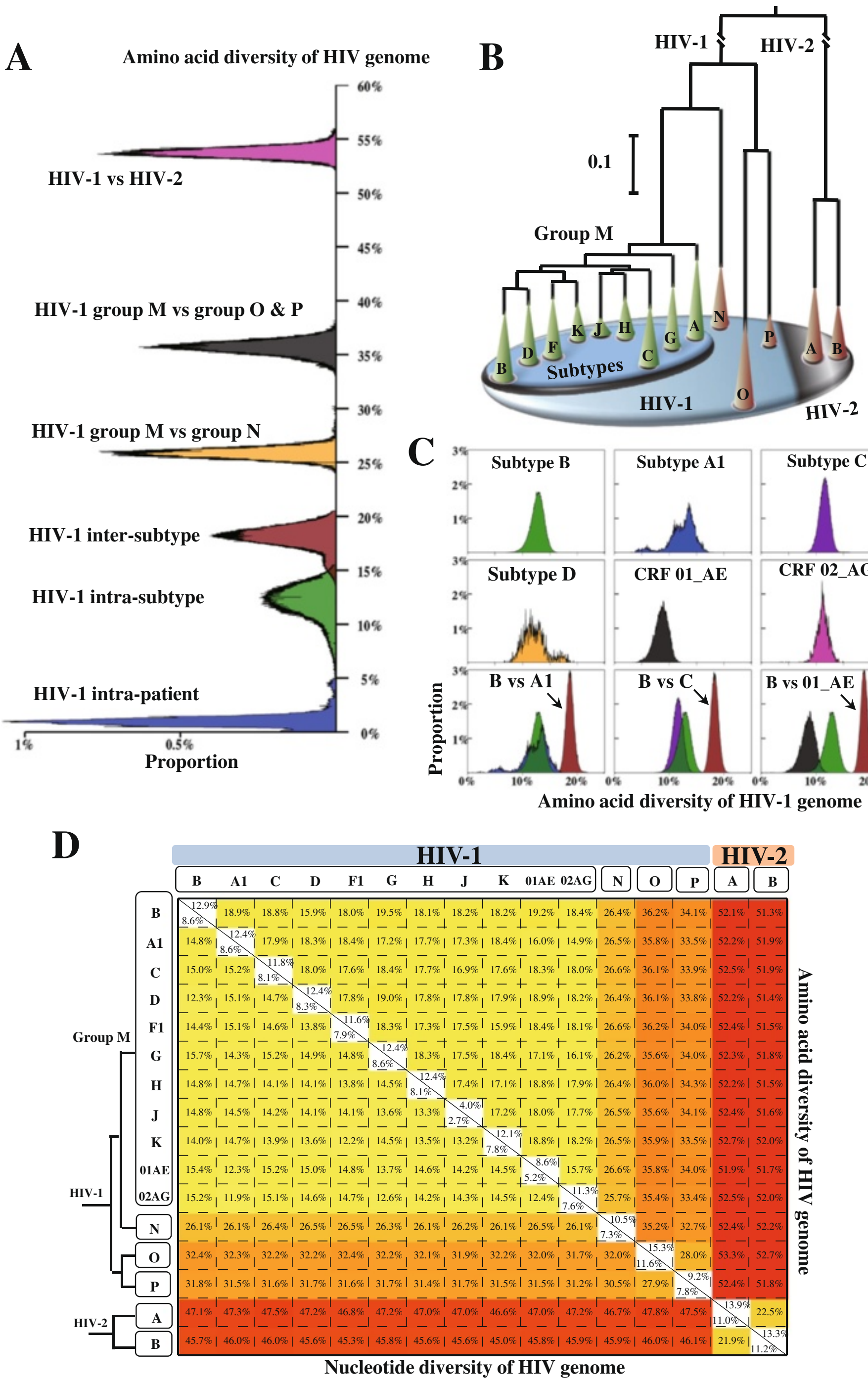
Genome-wide diversity within and across HIV types, major groups and subtypes

We quantified the nucleotide and amino acid diversity of the HIV genome using 2996 full-length sequences sampled from 1705 patients (Table 1). The amino acid diversity was 53.8% (95% confidence interval (CI): 53.0-54.6%) between HIV-1 and HIV-2, 41.1% (CI: 25.6-54.3%) between HIV-1 groups, 18.0% (CI: 15.6-19.6%) between HIV-1 subtypes, 12.0% (CI: 8.6-14.4%) within HIV-1 subtypes and 1.1% (CI: 0.3-2.2%) within HIV-1 patients (Figure 1A). Similarly, nucleotide genomic diversity was found to be the highest when comparing

Table 1 Information on HIV-1 and HIV-2 full-length genomic sequence datasets

Type	HIV-1														HIV-2	
Group	M											N	O	P	A	B
Subtype/CRF	A1	B	C	D	F1	G	H	J	K	01_AE	02_AG					
Number of genomes	159	1425	554	65	25	27	4	2	2	581	81	11	25	4	25	6
Number of patients	134	657	429	57	22	23	4	2	2	250	71	9	22	2	16	5
Average length in nucleotides*	8500	8600	8600	8500	8500	8600	8600	8600	8600	8500	8500	8500	8700	8600	8600	8600

*: Only the HIV coding regions are counted.



	HIV-1														HIV-2	
	B	A1	C	D	F1	G	H	J	K	01AE	02AG	N	O	P	A	B
B	12.9% 8.6%	18.9%	18.8%	15.9%	18.0%	19.5%	18.1%	18.2%	18.2%	19.2%	18.4%	26.4%	36.2%	34.1%	52.1%	51.3%
A1	14.8%	12.4% 8.6%	17.9%	18.3%	18.4%	17.2%	17.7%	17.3%	18.4%	16.0%	14.9%	26.5%	35.8%	33.5%	52.2%	51.9%
C	15.0%	15.2%	11.8% 8.1%	18.0%	17.6%	18.4%	17.7%	16.9%	17.6%	18.3%	18.0%	26.6%	36.1%	33.9%	52.5%	51.9%
D	12.3%	15.1%	14.7%	12.4% 8.3%	17.8%	19.0%	17.8%	17.8%	17.9%	18.9%	18.2%	26.4%	36.1%	33.8%	52.2%	51.4%
F1	14.4%	15.1%	14.6%	13.8%	11.6% 7.9%	18.3%	17.3%	17.5%	15.9%	18.4%	18.1%	26.6%	36.2%	34.0%	52.4%	51.5%
G	15.7%	14.3%	15.2%	14.9%	14.8%	12.4% 8.6%	18.3%	17.5%	18.4%	17.1%	16.1%	26.2%	35.6%	34.0%	52.3%	51.8%
H	14.8%	14.7%	14.1%	14.1%	13.8%	14.5%	12.4% 8.1%	17.4%	17.1%	18.8%	17.9%	26.4%	36.0%	34.3%	52.2%	51.5%
J	14.8%	14.5%	14.2%	14.1%	14.1%	13.6%	13.3%	4.0% 2.7%	17.2%	18.0%	17.7%	26.5%	35.6%	34.1%	52.4%	51.6%
K	14.0%	14.7%	13.9%	13.6%	12.2%	14.5%	13.5%	13.2%	12.1% 7.8%	18.8%	18.2%	26.5%	35.9%	33.5%	52.7%	52.0%
01AE	15.4%	12.3%	15.2%	15.0%	14.8%	13.7%	14.6%	14.2%	14.5%	8.6% 5.2%	15.7%	26.6%	35.8%	34.0%	51.9%	51.7%
02AG	15.2%	11.9%	15.1%	14.6%	14.7%	12.6%	14.2%	14.3%	14.5%	12.4%	11.3% 7.6%	25.7%	35.4%	33.4%	52.5%	52.0%
N	26.1%	26.1%	26.4%	26.5%	26.5%	26.3%	26.1%	26.2%	26.1%	26.5%	26.1%	10.5% 7.3%	35.2%	32.7%	52.4%	52.2%
O	32.4%	32.3%	32.2%	32.2%	32.4%	32.2%	32.1%	31.9%	32.2%	32.0%	31.7%	32.0%	15.3% 11.6%	28.0%	53.3%	52.7%
P	31.8%	31.5%	31.6%	31.7%	31.6%	31.7%	31.4%	31.7%	31.5%	31.5%	31.2%	30.5%	27.9%	9.2% 7.8%	52.4%	51.8%
A	47.1%	47.3%	47.5%	47.2%	46.8%	47.2%	47.0%	47.0%	46.6%	47.0%	47.2%	46.7%	47.8%	47.5%	13.9% 11.0%	22.5%
B	45.7%	46.0%	46.0%	45.6%	45.3%	45.8%	45.6%	45.6%	45.0%	45.8%	45.9%	45.9%	46.0%	46.1%	21.9%	13.3% 11.2%

Figure 1 (See legend on next page.)

(See figure on previous page.)

Figure 1 Distribution of HIV genome-wide diversity and phylogenetic tree. (A) Distribution plots of amino acid diversity in the HIV genome. The plots show the genomic diversity within HIV-1 infected patients (HIV-1 intra-patient, blue), within HIV-1 subtypes (HIV-1 intra-subtype, green), between HIV-1 subtypes (HIV-1 inter-subtype, red), between HIV-1 group M and group N (HIV-1 inter-group, yellow), between HIV-1 group M and group O/P (HIV-1 inter-group, black) and between HIV-1 and HIV-2 (pink). Distribution plots of nucleotide genomic diversity are shown in Additional file 1: Figure S2. **(B)** Maximum likelihood phylogenetic tree of HIV groups and pure subtypes. Green cones indicate HIV-1 subtypes in group M, while orange cones denote other HIV groups. All phylogenetic branches have bootstrap supports of more than 85% except one containing subtypes J, H and C. Branch lengths from the root to HIV-1 and HIV-2 are shortened for visualization purposes. SIV strains were not included in our phylogenetic tree. Visualization software: FigTree V1.4.0 (http://tree.bio.ed.ac.uk/software/figtree/). **(C)** Distribution plots of amino acid diversity in 6 major HIV-1 subtypes and CRFs (B, A1, C, D, CRF01_AE, CRF02_AG). X- and y-axes indicate the amino acid diversity and the proportions of sequence pairs, respectively. Six subplots in the first and second rows show the intra-subtype amino acid diversity of 6 HIV-1 subtypes and CRFs. Three subplots in the third row show the distribution of inter-subtype genomic diversity (B vs A1, B vs C, B vs 01_AE). One genomic sequence per patient (Table 1) was used for our analysis. Distribution plots of the other inter-clade genomic diversity are shown in Additional file 1: Figure S3. **(D)** Average inter- and intra-clade genomic diversity of HIV-1 and HIV-2. The top right matrix demonstrates results for amino acid diversity, the bottom left matrix for nucleotide diversity. HIV subtypes and groups are shown on the left side of the matrix.

HIV-1 and HIV-2 (mean: 48.32%, CI: 47.8-48.9%), followed by HIV-1 inter-group (37.5%, CI: 26.0-45.7%), HIV-1 inter-subtype (14.7%, CI: 12.2-15.8%), HIV-1 intra-subtype (8.2%, CI: 5.3-10.0%) and HIV-1 intra-patient diversity (0.6%, CI: 0.2-1.4%) (Additional file 1: Figure S2). As expected, the trend in HIV genomic diversity corresponds with the phylogenetic relationships between groups and pure subtypes in HIV-1 and HIV-2 (Figure 1B).

We next quantified genomic diversity within and between individual HIV clades (Figure 1C, Additional file 1: Figure S3). Within each HIV clade, amino acid diversity was consistently higher than nucleotide diversity (Figure 1D). CRF01_AE showed the lowest genomic diversity (nucleotide: 5.7%, amino acid: 8.7%) among the 10 HIV-1 subtypes with at least 10 sequences available (Figure 1D). Sequence variability was not uniformly distributed along the full-length HIV genome, but similar patterns were consistently observed in HIV group, subtype and CRF genomes at the nucleotide and amino acid levels (Figure 2A, B). Moreover, the estimated geographical distribution of HIV-1 genomic diversity (Additional file 1: Figure S4) showed a good agreement with the reported geographical distribution of HIV-1 subtypes [29].

Among all HIV proteins, Integrase was the most conserved protein (mean ± deviation: 4.5 ± 1.1%), while GP120 varied the most (21.3 ± 2.5%) (Table 2). Pairwise comparisons of genetic diversity between subtype B and the other clades identified conserved regions in the Capsid, Nucleocapsid, Protease, RT, Integrase, Vpr and the N terminus of GP41 (Figure 2C). Despite the different degrees of sequence diversity along the full-length genome, the nucleotide and amino acid compositions were comparable across the 16 group and subtype genomes (Figure 3A, 3B).

Multiple factors shape HIV-1 genomic diversity

We next evaluated three potential factors (protein multimerization, immunological constraints, HIV-human protein interactions) that shaped the HIV genomic diversity. Firstly, we calculated the average diversity at amino acid positions of the 15 HIV-1 proteins (Figure 3C). For every HIV-1 group, subtype and CRF, the average amino acid diversity was significantly higher in the monomeric proteins (Nucleocapsid, Vpr, Vpu, p6) than in the multimeric proteins (Matrix, Capsid, Protease, RT, Integrase, Vif, Tat, Rev, GP120, GP41, Nef) ($p\text{-value} < 0.05$) (Additional file 2: Table S1). This suggests that protein multimerization imposes a constraint on the HIV-1 amino acid variability. Besides our diversity analysis, we also measured synonymous (dS) and non-synonymous (dN) substitution rates in the HIV-1 subtype A1, B, C and CRF01_AE genomes, which contained more than 100 sequences in our genomic datasets (Additional file 2: Table S2). Mann-Whitney U tests were used to compare the distribution of dN, dS and the ratio of non-synonymous to synonymous substitutions (dN/dS) between the monomeric and multimeric protein groups. We found that the distributions of dN/dS and dN in these two groups were significantly different ($p\text{-value} < 0.05$) in HIV-1 subtypes A1, B, C and CRF01_AE (Additional file 2: Table S3). This suggests that multimeric proteins are under the stronger negative selective pressure than monomeric proteins.

Secondly, we evaluated the amino acid variation in the known CD4 T cell, CD8 T cell and antibody epitopes (see Materials). By measuring the diversity of 3066 amino acid positions, we identified 919 (30%) variable positions with amino acid diversity above 12.9% (the average amino acid diversity within subtype B) using 657 subtype B genomic sequences. Univariate analysis showed that these variable positions were preferably located within antibody epitopes (OR 1.43, CI: 1.15-1.79, Fisher's exact test, $p\text{-value} = 0.0015$) and CD4 T cell epitopes (OR 1.73, CI: 1.18-2.96, $p\text{-value} = 0.0438$), but not within CD8 T cell epitopes (OR 1.11, CI: 0.82-1.51, $p\text{-value} = 0.498$) (Figure 2A).

Thirdly, we mapped 1352 interactions between 1052 human and 15 HIV-1 proteins using the HIV-human protein interaction dataset (Figure 3D, see Materials). The following three observations support the hypothesis

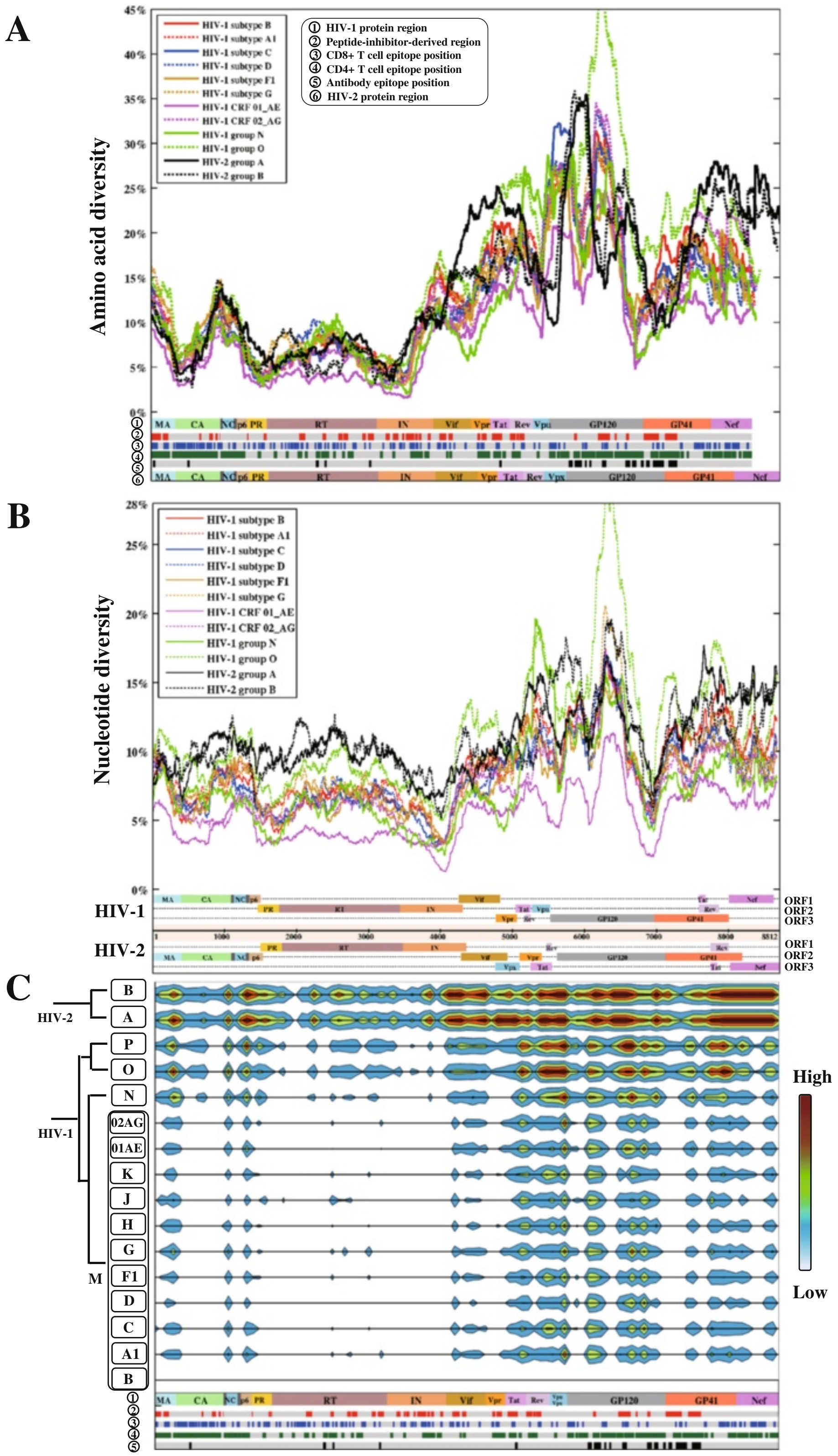


Figure 2 (See legend on next page.)

(See figure on previous page.)

Figure 2 Plots of amino acid and nucleotide diversity in the HIV full-length genome. (A) Amino acid diversity along the HIV full-length genome using the sliding windows (window size: 100AA; also see the plots of exact diversity values in Additional file 1: Figure S5). Each colored plot shows the density of amino acid diversity for one HIV group, subtype or CRF genome, indicated by the figure legend. Six layers are shown beneath the plots: (1) HIV-1 protein regions (HXB2 reference) are concatenated and shown with abbreviated names (e.g. MA: matrix); (2) peptide-inhibitor-derived region; (3) CD8+ T cell epitope position; (4) CD4+ T cell epitope position; (5) antibody epitope position; (6) HIV-2 protein region (BEN reference). **(B)** Nucleotide diversity along the full-length HIV genome using sliding windows (window size: 300 nucleotides; also see the plots of exact diversity values in Additional file 1: Figure S6). Each colored plot shows the density of nucleotide diversity for one HIV group, subtype or CRF genome, indicated by the figure legend. Annotated HIV-1 and HIV-2 reference genomes are shown beneath; each track contains one open reading frame (ORF). Long terminal regions in the HIV genome are not shown. **(C)** Contour map of inter-clade amino acid diversity between HIV-1 subtype B and the other HIV genomes. Inter-clade amino acid diversity was calculated by a sliding window of 30 amino acids over the HIV genome (low: ≤1 AA difference, high: ≥25 AA differences). Five colored layers beneath the contour map are annotated similarly in **(A)**.

that the amino acid diversity of HIV-1 proteins is associated with HIV-human protein interactions. (1) Univariate analysis showed that HIV-1 proteins with higher amino acid diversity interact with more human proteins (Pearson's coefficient = 0.74, p-value = 0.0017). Polynomial regression analysis further identified a second-order model that fitted the correlation between these two variables (Figure 4A, adjusted R-squared: 0.82). (2) Intrinsically disordered structures in HIV-1 proteins can interact with multiple interaction partners [30]. Univariate analysis showed a significant correlation between the average amino acid diversity and the average disorder scores of HIV-1 proteins (Pearson's coefficient = 0.64, p-value = 0.015, Figure 4B). (3) The levels of HIV-human protein interactions clustered according to the functional roles of the HIV-1 proteins, which have different functional roles and requirements for interactions with human proteins (Figure 4C). HIV regulatory proteins (Tat, Rev) and envelope proteins (GP120, GP41) had the largest number of interactions with different human proteins (568 for the regulatory proteins, 322 for the envelope proteins), while viral enzymes had the least number of interactions (Figure 4C). The average amino acid diversity of envelope proteins (20.4%) and regulatory proteins (18.8%) was higher than that of accessory proteins (16.0%), structural proteins (9.0%) and viral enzymes (5.9%) (Additional file 1: Figure S7). Our findings suggest that HIV-1 proteins with higher genetic diversities have larger intrinsically disordered structures and interact with more human proteins.

Peptide inhibitors are mainly derived from conserved subtype B genomic regions

We investigated the 121 HIV-derived peptide inhibitors reported between 1993 and 2013 (Additional file 2: Table S4). Figure 5A illustrates the GP41 structure and the GP41-derived region of T20 as an example of HIV-derived peptide inhibitors. Peptide inhibitors had on average a length of 25 AAs (range: 3 to 73), a charge of +0.27 at pH 7.2 and a molecular weight of 2953 g/mol. Most common amino acids in these peptide inhibitors were leucine, glutamic acid and isoleucine (Additional file 1: Figure S8). Comparisons

Table 2 Average amino acid diversity of viral proteins within individual HIV clades (%)

	Clade #	MA	CA	NC	p6	PR	RT	IN	Vif	Vpr	Tat	Rev	Vpu/Vpx*	GP120	GP41	Nef
HIV-1	**Subtype A1**	13.07	7.3	7	19.89	5.59	6.44	4.76	13.63	9.86	18.14	16.83	23.04	23.34	13.38	14.64
	Subtype B	12.92	4.95	10.83	14.99	8.21	6.04	4.91	14.71	11.29	19.32	18.34	20.04	23.93	15.59	17.79
	Subtype C	16.11	5.77	9.39	15.52	6.15	5.84	4.45	10.61	10.64	15.37	16.10	20.21	22.89	14.19	14.96
	Subtype D	14.78	5.08	10.85	13.16	7.68	6.98	4.80	11.57	11.74	17.89	15.87	19.40	23.58	13.61	14.42
	Subtype F1	13.69	5.71	10.07	16.72	8.84	6.42	4.76	11.04	9.26	17.01	15.70	18.27	20.24	13.39	14.02
	Subtype G	17.33	5.54	7.40	17.21	6.60	7.30	4.17	14.36	10.33	15.51	18.39	19.25	23.87	12.98	14.15
	Subtype H	18.81	3.61	8.48	17.65	4.55	5.95	5.71	15.46	9.11	19.28	19.80	18.31	21.51	12.33	16.34
	CRF01_AE	10.06	3.00	4.92	13.37	4.34	3.75	2.39	9.32	7.97	13.57	13.34	12.80	17.40	9.62	11.78
	CRF02_AG	13.57	5.69	3.78	13.63	5.56	5.67	4.06	13.44	7.70	13.58	15.49	16.42	23.37	12.09	15.45
	Group N	10.52	6.40	7.06	10.01	5.93	6.78	3.09	11.02	4.87	12.15	12.77	35.01	21.43	9.15	12.64
	Group O	14.52	7.45	9.89	22.17	8.20	6.51	5.77	16.17	12.16	21.61	21.25	27.12	28.65	19.35	17.96
HIV-2	**Group A**	10.15	4.67	11.82	11.02	7.14	6.89	6.78	12.74	22.55	22.59	22.75	10.40	18.55	16.47	21.98
	Group B	12.14	3.54	13.15	15.4	5.76	6.4	6.19	12.71	12.66	20.44	15.82	11.45	21.20	15.29	20.59

#: Only HIV groups, subtypes or CRFs with more than 2 genomic sequences are listed.
*: Vpu in HIV-1 and Vpx in HIV-2.

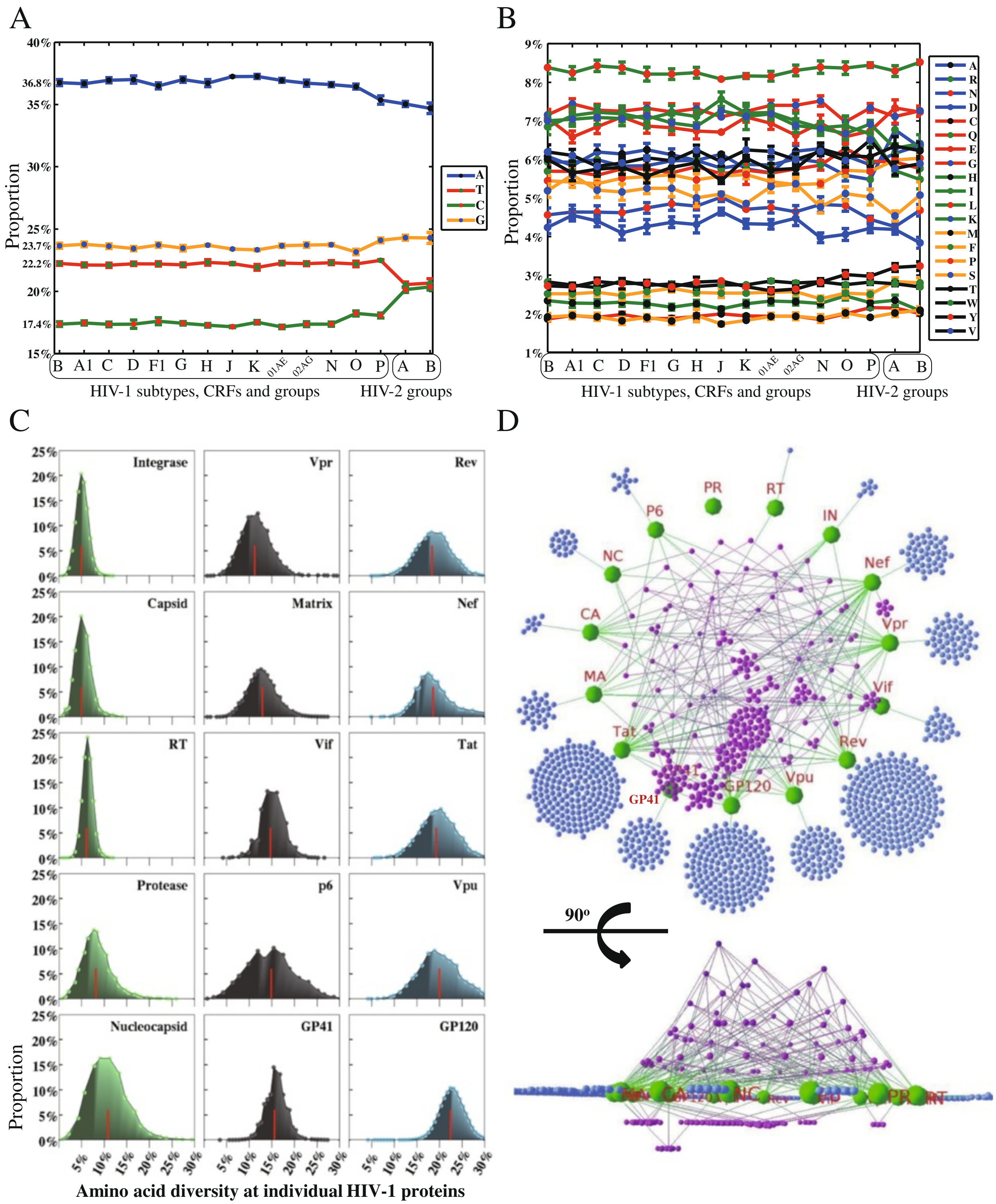


Figure 3 (See legend on next page.)

(See figure on previous page.)

Figure 3 Nucleotide and amino acid composition of HIV genomes and 3D mapping of HIV-human protein interactions. (A) Nucleotide composition for HIV-1 and HIV-2. X-axis represents the HIV groups, subtypes and CRFs. Y-axis shows the average proportions of nucleotides (A, T, C, G) using the HIV genomic sequence datasets (one sequence per patient, Table 1). **(B)** Amino acid composition for HIV-1 and HIV-2. X-axis represents HIV groups, subtypes and CRFs. Y-axis shows the average proportions of amino acids using the HIV protein sequence datasets (one sequence per patient, Table 1). **(C)** Distribution plots of amino acid genetic diversity for 15 HIV-1 subtype B proteins. Each subplot demonstrates a viral protein. X- and y-axes indicate the amino acid diversity and the proportions of amino acid diversity, respectively. Red lines inside the distribution plots indicate the mean values of amino acid diversity at individual proteins. **(D)** Top and side views of 3D HIV-human protein interaction networks. HIV-1 proteins with protein names annotated are indicated by green spheres. Human proteins that interact with only one HIV-1 protein are indicated by blue spheres in the outer circle (one human protein one sphere). Human proteins that interact with more than one HIV-1 protein are indicated by purple spheres above the plane of HIV-1 proteins. The height of the layers above the plane indicates the number of HIV proteins that a human protein interacts with. Below, human proteins are clustered if they interact with a set of more than one HIV-1 protein. Abbreviation names have been described in the abbreviation list. Visualization software: Geomi V2.0(http://sydney.edu.au/engineering/it/~visual/geomi2/).

between the 121 peptide sequences and the consensus sequences of 16 HIV group, subtype and CRF genomes showed the highest sequence similarity with subtype B (79.8%) (Figure 5B). Aspartic acid to asparagine (25.7%) was the most common amino acid substitution between the consensus subtype B sequence and the peptide inhibitor sequences (Figure 5C).

We characterized peptide-derived regions in the subtype B genome. Of the 894 amino acid positions from which the 121 peptide inhibitors were derived, 41.2% were located in helix structures and 60.2% displayed less than 5% genetic diversity in the subtype B genome. Forty-two inhibitors had IC_{50} or EC_{50} values less than 1 μM and were derived from 249 amino acid positions in the HIV-1 genome (Additional file 2: Table S4). In the subtype B genome, these 249 positions displayed significantly lower amino acid diversity compared to the genome-wide diversity (Figure 5D, 10.1% vs 12.9%, p-value = 0.019), and were likely to be from conserved (amino acid diversity < 5%, OR: 1.43 (1.09-1.88), p-value = 0.016), solvent exposed (OR: 2.47 (1.88-3.24), p-value = 3.9E-11) and intrinsically ordered structures (disorder score < 0.4, OR: 1.75 (1.21-2.51), p-value = 0.0019) (Figure 5E).

Integrated findings from our analyses on HIV-1 genomic diversity, HIV-derived peptide inhibitors and protein structures are visualized in Figure 6. The HIV genomic sequence datasets and our toolbox developed for data visualization, genomic diversity analysis and HIV genomic alignments are freely available in Additional file 3.

Discussion and conclusions

To our knowledge, this study provides the first large-scale analysis that investigates the genomic variability of 16 major groups, subtypes and CRFs in HIV-1 and HIV-2. While previous studies have reported the diversity of HIV genomes in small cohorts of patients (n < 250) [11-24,31], our analyses evaluated HIV genome-wide diversity using 2996 full-length genomic sequences sampled from 1705 patients worldwide. We evaluated three important aspects of HIV genomic diversity using the integrated datasets of genomic sequences, protein structures, HIV-human protein interactions, human immune epitopes and HIV-derived peptide inhibitors. Firstly, we quantified HIV genomic diversity at the individual and population levels. Secondly, we reported possible associations between HIV-1 amino acid diversity and protein multimerization, immunological constraints and HIV-human protein interactions. Thirdly, we mapped conserved regions in the HIV genome and characterized experimental and clinically used HIV-derived peptide inhibitors [7].

Quantification of HIV genomic diversity

HIV-1 genomic diversity is the lowest within single patients and increases in the following order when different patients are considered: within subtypes, between subtypes, between groups and between HIV types (Figure 1). A nucleotide genomic diversity was quantified to be 48.3% between HIV-1 and HIV-2, 37.5% between HIV-1 groups, 14.7% between HIV-1 subtypes, 8.2% within HIV-1 subtypes, and 0.6% within single patients infected with HIV-1. These results are in a good agreement with previous studies which analyzed less than 100 sequences [13,23]. Our study quantified genomic diversity at the population level using the largest sequence dataset ever analyzed, thereby resulting in robust and accurate estimations.

As shown in Figure 2, the degree of HIV genetic diversity varied along the full-length genome. A comparison of the amino acid diversity of HIV proteins revealed the highest diversity in the envelope proteins, followed by the regulatory, accessory, structural and enzymatic proteins (Figure 4, Table 2). Our diversity analysis showed comparable results compared to our previous studies, which reported the amino acid diversity of Gag and Rev using 10862 and 4632 sequences respectively [9,32]. In contrast, estimated amino acid diversities for Pol (5.7 ± 0.9%) and Env (18.7 ± 2.7%) displayed higher values than

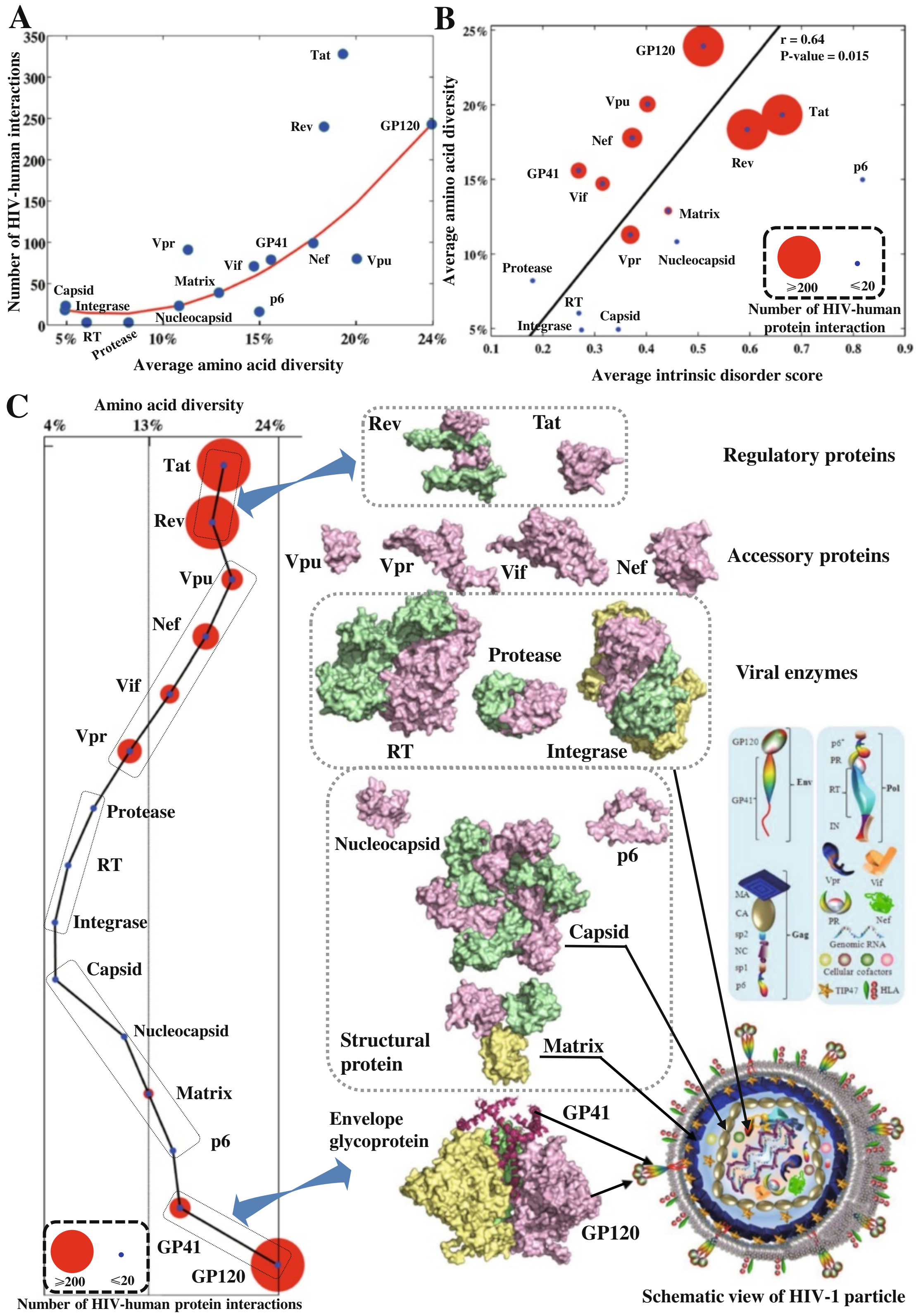


Figure 4 (See legend on next page.)

(See figure on previous page.)

Figure 4 Correlations between HIV-1 protein diversity and HIV-human protein interactions, protein disorder and viral particle structures. **(A)** Plot of polynomial regression between the HIV-1 protein diversity (x-axis) and the number of HIV-human protein interactions (y-axis). The second-order model is $\mathcal{Y} = 8346\mathcal{X}^2 - 1223\mathcal{X} + 57.96$ (adjusted R-squared: 0.82, root-mean-square error: 42.31). **(B)** Plot of average protein disorder score and average amino acid diversity in HIV-1 proteins. Red circles indicate the number of HIV-human protein interactions at individual viral proteins, for visualization purpose, scaled between 20 and 200 interactions (proteins with fewer than 20 interactions are scaled to the same size as those with 20, proteins with more than 200 interactions are scaled to the same size as those with 200). Average amino acid diversities of HIV-1 proteins are calculated using subtype B sequences (one genomic sequence per patient, Table 1). **(C)** Clustering of HIV-1 proteins and schematic view of HIV-1 viral particle. On the left, each colored circle represents a viral protein positioned according to the clusters of protein functions. The size of each red circle indicates the number of HIV-human protein interactions involving each HIV-1 protein (see **(B)**). On the right, the schematic view of mature viral particle is visualized at the bottom with annotations indicated in the inserted figure legend. Above, surface representations show the structures of HIV-1 proteins that are grouped according to their functional roles. Different units in HIV-1 multimeric proteins are indicated with different colors and HIV-1 monomeric proteins are colored pink. HIV-1 protein structures are scaled according to their precise protein sizes for direct comparison. Visualization software: PyMOL V1.5 (http://www.pymol.org/).

previous reports analyzing fewer than 100 sequences of subtypes A and B [24]. Using large-scale sequence datasets, our study thus provides a complete estimation of the genetic diversity in 16 HIV proteins.

HIV genomic diversity is shaped by multiple factors

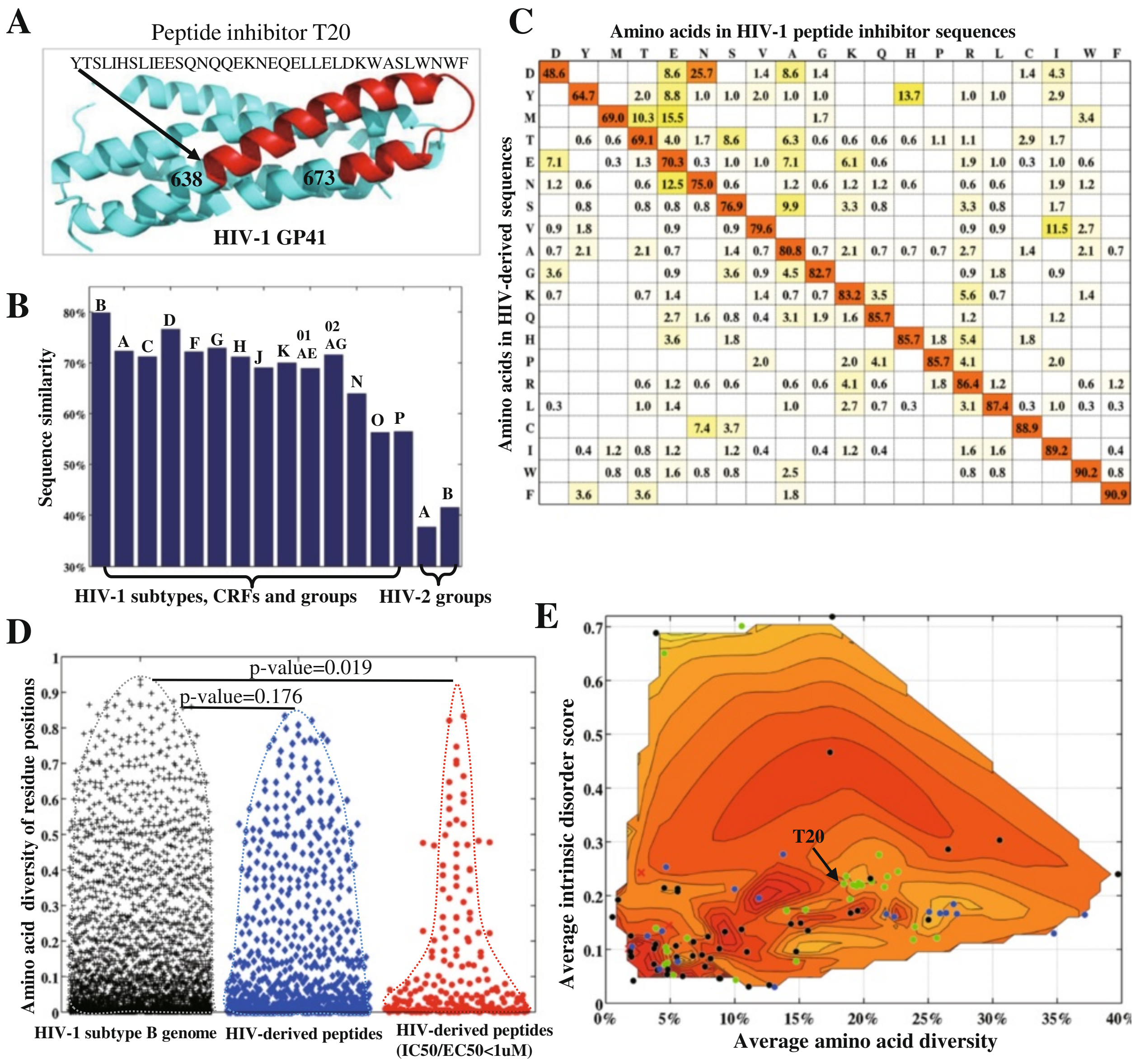
HIV genomic diversity is driven by the high rates of viral replication, recombination and mutation [33], but other factors also play a role in shaping HIV genomic diversity. For instance, positive selection was significantly identified in the overlapping reading frames of the HIV-1 subtype B genome, probably due to functional and structural requirements [26]. To evaluate other potential factors, we correlated HIV amino acid diversity with protein multimerization, human immunological constraints and HIV-human protein interactions. We showed that CD4 T cell and antibody epitope positions in the HIV-1 genome were likely to have high amino acid diversities, supporting the hypothesis that the human immune system imposes a diversifying selective pressure on the HIV-1 genome [26]. Moreover, we found that the average amino acid diversity was significantly lower in the multimeric than in the monomeric proteins, suggesting that protein multimerization places a constraint on HIV-1 sequence conservation. While synonymous substitution rates were not significantly different, nonsynonymous substitution rates were significantly lower in multimeric proteins (Additional file 2: Table S3), suggesting that the negative selective pressure may be stronger for multimeric proteins. Previous findings on other protein families have also shown that multimeric proteins are relatively conserved and have less tolerance for amino acid substitutions, probably because of the structural and functional constraints [34-37].

We mapped 1352 HIV-human protein interactions between 15 HIV-1 proteins and 1052 human proteins. A strong association was found between the amino acid diversity of HIV-1 proteins and the number of HIV-human protein interactions (Figure 4A). HIV-1 proteins with higher genetic diversities tended to interact with more human proteins. This is probably associated with structurally disordered regions in HIV-1 proteins (Figure 4B), which provide the structural flexibility for HIV to interact with multiple human proteins [30]. For instance, GP120 uses five hypervariable loops (Additional file 1: Figure S12) to interact with various human proteins [38]. An intricate landscape of HIV-human protein complexes is made by HIV to exploit human cellular machineries during the HIV infection and production [39]. Despite the high variability of HIV, it is surprising that the nucleotide and amino acid compositions were remarkably constant across all HIV-1 and HIV-2 clades (Figure 3A, 3B), suggesting that other constraints may be active to restrict the HIV genetic diversity [25]. For instance, HIV RNA structures require the stable nucleotide compositions for the reverse transcription [25].

Conserved drug targets in the HIV-1 genome

Many peptide inhibitors derived from HIV-1 proteins have shown promising antiviral activities and some of these inhibitors are currently under clinical trials [40,41]. Our study summarized HIV-derived peptide inhibitors published between 1993 and 2013 (Additional file 1: Figure S12-S21, Additional file 2: Table S4), and mapped the positions of these inhibitors to the HIV-1 genome (Figure 6). We showed that most peptide inhibitors were derived from the regions of HIV-1 subtype B proteins (Figure 5B), which had conserved, solvent exposed and intrinsically ordered structures (Figure 5E). This information enhances current understanding of HIV-derived peptide inhibitors, which may provide valuable guidelines for the design of novel peptide inhibitors [42,43].

In the full-length genome, we identified conserved regions in Capsid, Nucleocapsid, Protease, RT, Integrase, Vpr and N-terminal domain of GP41 (Figure 2). These conserved regions have been targeted by known anti-HIV inhibitors (Figure 6). For instance, over 40 experimental inhibitors with promising antiviral activities have been designed to target Capsid and Nucleocapsid [9]. HIV enzymes (Protease, RT, Integrase) are targeted by the majority of the FDA-approved



Amino acids in HIV-1 peptide inhibitor sequences

Amino acids in HIV-derived sequences

	D	Y	M	T	E	N	S	V	A	G	K	Q	H	P	R	L	C	I	W	F
D	48.6				8.6	25.7		1.4	8.6	1.4							1.4	4.3		
Y		64.7		2.0	8.8	1.0	1.0	2.0	1.0	1.0			13.7		1.0	1.0		2.9		
M			69.0	10.3	15.5					1.7									3.4	
T		0.6	0.6	69.1	4.0	1.7	8.6		6.3	0.6	0.6	0.6	0.6	1.1	1.1		2.9	1.7		
E	7.1		0.3	1.3	70.3	0.3	1.0	1.0	7.1		6.1	0.6			1.9	1.0	0.3	1.0	0.6	
N	1.2	0.6		0.6	12.5	75.0	0.6		1.2	0.6	1.2	1.2	0.6		0.6	0.6		1.9	1.2	
S		0.8		0.8	0.8	0.8	76.9		9.9		3.3	0.8			3.3	0.8		1.7		
V	0.9	1.8			0.9		0.9	79.6							0.9	0.9		11.5	2.7	
A	0.7	2.1		2.1	0.7		1.4	0.7	80.8	0.7	2.1	0.7	0.7	0.7	2.7		1.4		2.1	0.7
G	3.6				0.9		3.6	0.9	4.5	82.7					0.9	1.8		0.9		
K	0.7			0.7	1.4			1.4	0.7	0.7	83.2	3.5			5.6	0.7			1.4	
Q					2.7	1.6	0.8	0.4	3.1	1.9	1.6	85.7			1.2			1.2		
H					3.6		1.8						85.7	1.8	5.4		1.8			
P								2.0			2.0	4.1		85.7	4.1			2.0		
R				0.6	1.2	0.6	0.6		0.6	0.6	4.1	0.6		1.8	86.4	1.2			0.6	1.2
L	0.3			1.0	1.4				1.0		2.7	0.7	0.3		3.1	87.4	0.3	1.0	0.3	0.3
C						7.4	3.7										88.9			
I		0.4	1.2	0.8	1.2		1.2	0.4		0.4	1.2	0.4			1.6	1.6		89.2		0.4
W			0.8	0.8	1.6	0.8	0.8		2.5						0.8	0.8			90.2	0.8
F		3.6		3.6					1.8											90.9

Figure 5 Characterization of HIV-derived peptide inhibitors. (A) Cartoon representation of GP41 structure. The red structure indicates the region from which peptide inhibitor T20 was derived (PDB: 3H01). **(B)** Bar plot of sequence similarities between peptide inhibitor sequences and the sequences of HIV-derived regions in the consensus genome of different HIV clades. X-axis presents the HIV groups, subtypes and CRFs. Y-axis shows the sequence similarity between peptide inhibitor sequences and the sequences of HIV-derived regions in the consensus genomes of HIV groups, subtypes or CRFs. **(C)** Amino acid replacements between peptide inhibitor sequences and HIV-derived regions in the subtype B genome. The percentage values (%) are colored using heat maps. **(D)** Distribution (bee-swarm) plots of amino acid diversity in the full-length subtype B genome (black crosses), peptide-derived regions (blue diamonds) and peptide-derived regions of those inhibitors whose IC_{50}/EC_{50} are less than 1 μM (red circles). Each shape represents the amino acid diversity at one protein position. Two-sample Kolmogorov-Smirnov tests were performed to compare diversity distributions (significance level: 0.05). **(E)** Plot of amino acid diversity (x-axis), disorder score (y-axis) and solvent accessible surface area of peptide-inhibitor-derived regions (contour map, darker red indicates larger accessible surface areas). GP41 inhibitor T20 is also annotated. For individual peptide inhibitors, the average amino acid diversity, disorder score and solvent accessible surface areas are shown in Additional file 1: Figure S9, S10 and S11, respectively.

antiretroviral drugs. Peptide inhibitor T20 targets the N-terminal heptad domain of GP41 [44]. Overall, our sequence analysis mapped the conserved drug target regions in the HIV-1 genome, providing useful information for drug design.

Implications for HIV vaccine development

HIV subtype- and geography-specific vaccination has been proposed to contend with the challenges imposed by the high HIV genetic diversity [2]. Previous vaccine trials were carried out in regional populations dominated by a single

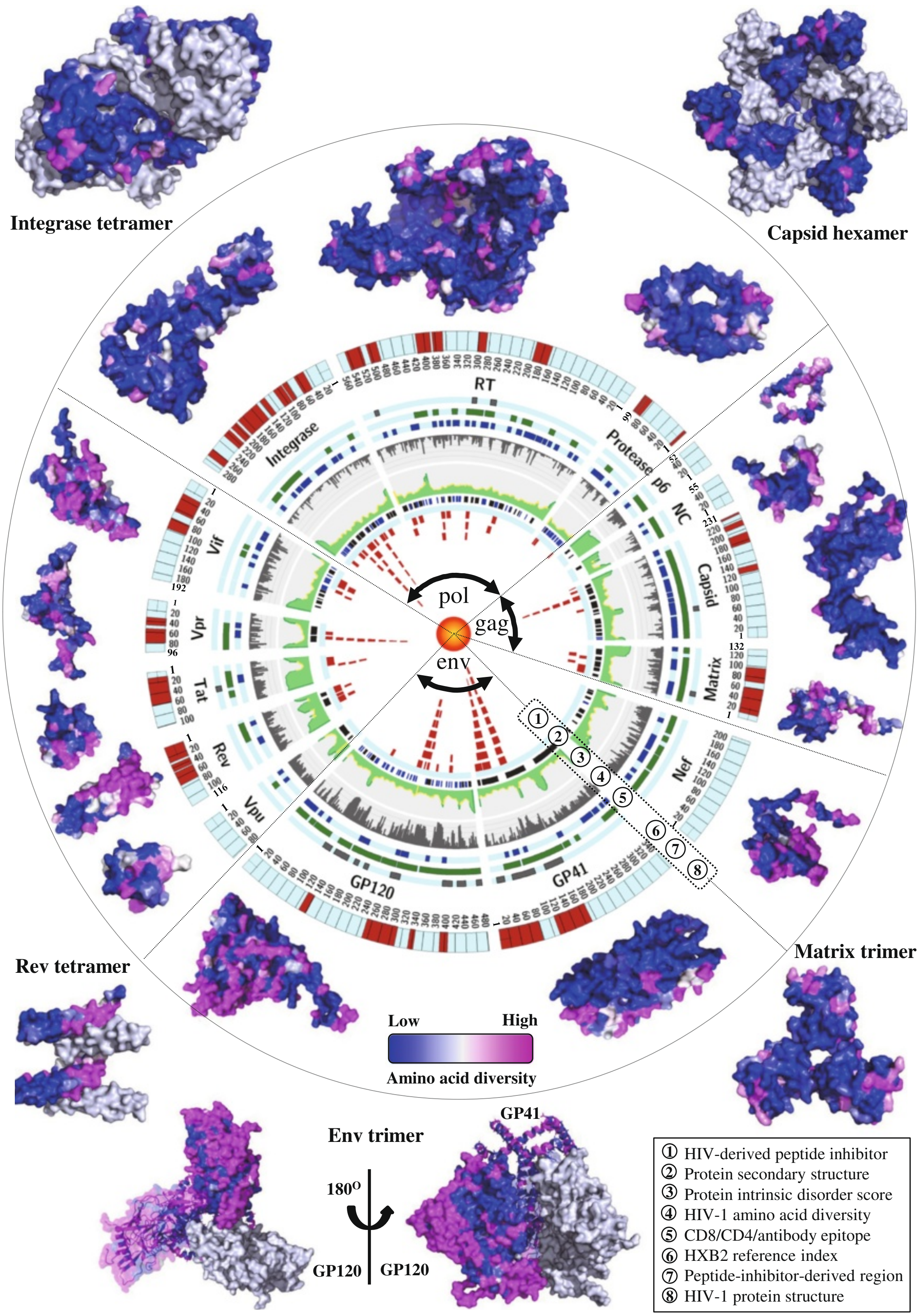

Integrase tetramer
Capsid hexamer
RT
Protease
p6
NC
Capsid
Matrix
Nef
GP41
GP120
Vpu
Rev
Tat
Vpr
Vif
Integrase
pol
gag
env
Matrix trimer
Rev tetramer
Low
High
Amino acid diversity
Env trimer
GP41
180°
GP120
GP120
① HIV-derived peptide inhibitor
② Protein secondary structure
③ Protein intrinsic disorder score
④ HIV-1 amino acid diversity
⑤ CD8/CD4/antibody epitope
⑥ HXB2 reference index
⑦ Peptide-inhibitor-derived region
⑧ HIV-1 protein structure

Figure 6 (See legend on next page.)

(See figure on previous page.)

Figure 6 An integrated map of HIV-1 genomic diversity and protein structures. All 15 HIV-1 proteins are mapped in the circle with 8 layers, showing the schematic view of HIV-1 peptide inhibitors (layer 1), protein secondary structures (layer 2, dark blue: helix structures, light blue: beta-strand structures, white: random-coil structures), protein disorder scores (layer 3), amino acid diversity at individual positions (layer 4), human CD4/CD8/antibody epitope regions (layer 5, three sub-layers from inside to outside represent CD8+ T cell, CD4+ T cell and antibody epitope regions), HXB2 reference indices (layer 6), peptide-inhibitor-derived regions (layer 7) and the protein structures are colored according to the diversity of amino acid positions (layer 8, low: 0%, high: ≥30%). Three major genes (*gag*, *pol*, *env*) are annotated in the center. Structures of multimeric HIV-1 proteins are shown outside the circle and different protein units are colored separately. The list of PDB data is available in Additional file 2: Table S5. Visualization software: Circos V0.64 (http://circos.ca/).

HIV-1 subtype or CRF. For instance, the STEP [11] and RV144 [12] vaccine trials targeted patient populations mainly infected by subtype B and CRF01_AE, respectively. Particularly, the RV144 trial in 2009 showed the first sign that a prime–boost strategy achieved a modest vaccine efficacy (31.2%) in the heterosexual population, which was at risk for infections with CRF01_AE [12,31]. In our analysis, CRF01_AE has the lowest genomic diversity among the 12 analyzed HIV groups, subtypes and CRFs (Figures 1D, 2A, 2B). It is thus tempting to speculate that the low diversity of CRF01_AE may have contributed to the success of the RV144 trial.

As conserved epitopes are ideal targets for potential vaccines to contend with the high HIV diversity [24,45], our study highlighted position-specific conservation along the full-length HIV genome (Figure 6). Moreover, HIV-1 consensus sequences have been considered as potential vaccine candidates to minimize genetic diversity between vaccine candidates and circulating strains [2]. Previous analyses on fewer than 100 Matrix and GP160 sequences reported that genetic diversity between subtype-specific consensus sequences and circulating strains was only half of the genetic diversity between circulating strains from the same subtype [2]. We found that in the full-length HIV genome, this effect was much smaller as we only observed a 32.5% reduction of the genomic diversity (8.3% vs. 12.3%, Additional file 1: Figure S22). As the most explored vaccine target protein, GP120 has the highest genetic diversity among all HIV proteins (Table 2), presenting a challenge in the search for broadly neutralizing antibodies and vaccines [46]. Furthermore, we mapped the global distribution of HIV-1 genomic diversity (Additional file 1: Figure S4). Our results showed the highest HIV genomic diversity in Central Africa, the birthplace of HIV [1,29], which suggested the difficulty of implementing HIV vaccines in this region.

Limitations and future perspectives

The limited number of full-length genomic sequences for HIV-1 subtypes H, J and K, group P and HIV-2 group B (n < 10) may have affected our estimation of sequence diversity, but consistent patterns were detected in the full-length genome across different HIV groups and subtypes (Figure 2). Sequence diversity investigations at a population level for such clades may improve when adding more sequences from individual proteins. However, it would also result in difficulties (e.g. lack of patient or treatment information) that may affect our confidence in data quality, and consequently the accuracy of our results. Our structural analysis focuses on HIV-1 proteins because most PDB data are available for HIV-1 but not for HIV-2 (Additional file 2: Table S5). Beside the multiple factors described in our study, other driving forces may shape HIV genetic diversity [25,47] and the genetic diversity data reported in our study can be useful for further investigations. Information on positions involved in HIV-human protein interactions is largely lacking, restricting our analysis from exploring the genetic diversity of protein interaction positions. Despite an extensive search, anti-HIV peptide inhibitors other than the ones described here may have been developed, but major changes in our conclusions regarding the known peptide-derived regions are not expected. Future studies are still needed to clarify how to improve vaccines and anti-HIV inhibitors based on the information of HIV genomic diversity. The increased knowledge of genome-wide diversity from our study may contribute to a better rational design of HIV vaccines and inhibitors.

Methods

HIV genomic sequence dataset

In August 2013, we retrieved 3607 nucleotide genomic sequences of major HIV-1 and HIV-2 clades (HIV-2 group A and B, HIV-1 group N, O, P, subtype A1, B, C, D, F1, G, H, J, K, CRF01_AE, CRF02_AG) from the HIV Los Alamos database (www.hiv.lanl.gov/). HIV-1 subtype was determined by the Rega [48] and COMET [49] subtyping tools. The quality criteria for removing duplicates and sequences with hypermutations, stop codons, ambiguous nucleotides or discordant subtype classification were described in [9]. The sequence dataset that fulfilled the quality criteria comprised 2996 genomic sequences, sampled from 1684 HIV-1 and 21 HIV-2 patients between 1982 and 2013. Information on genomic sequence datasets is summarized in Table 1.

Nucleotide genomic sequences were aligned using MUSCLE [50]. Protein regions encoded by their respective open reading frames (ORFs) were concatenated according to the reference strains (HIV-1: HXB2, HIV-2: BEN). For each HIV protein coding region, the translation of nucleotide to amino acid sequence alignments was optimized by our nucleotide to amino acid alignment toolbox. This toolbox maximizes amino acid matches, including in overlapping reading frames, based on the BLOSUM62 substitution matrix. Sequence alignments were further curated using Seaview v4.3 [51]. To show the alignment quality, we measured the percentages of deletions and insertions in the multiple sequence alignments (MSAs) of HIV genomes, which were less than 1.31% in 16 individual HIV clades (Additional file 2: Table S6). Our alignment toolbox and genomic sequences are available in Additional file 3.

PDB, HIV-human protein interaction, CD4/CD8/antibody epitope datasets

As of February 2014, we queried HIV PDB data from the RCSB Protein Data Bank using sequence search; PDB quality was then examined using PDBREPORT [52] (Additional file 2: Table S5). We extracted HIV-human protein interactions (interaction type: physical interaction) from the NCBI HIV-1 human protein interaction database [53]. From the HIV Los Alamos database (http://www.hiv.lanl.gov/content/immunology/), we extracted the human CD4 T cell and antibody epitopes in HIV-1. For human CTL/CD8 T cell epitopes, we included the best-defined CTL epitopes of the A-list described in [54] (Additional file 2: Table S7).

HIV-derived peptide inhibitor dataset

HIV-derived peptide inhibitors have their amino acid sequences derived from HIV proteins. We searched for English articles in PubMed published between January 1983 and September 2013 using the keywords “HIV peptide inhibitor”, “HIV [protein name] peptide” and “HIV [protein name] inhibitor”. References from primary studies, review articles and peptide design papers were also reviewed. If more than one peptide inhibitor were reported in one publication, only the most promising peptide inhibitors as indicated by the abstract of articles were collected. If data on the same inhibitors was reported by more than one publication, only the latest results were retained. Additional file 2: Table S4 summarizes the 121 peptide inhibitors with corresponding information on peptide sequences, peptide-derived regions, target proteins, inhibitory activities and references.

Protein secondary structure

For HIV-1 proteins (Rev, GP41) whose crystalized structures are not fully resolved in the PDB data, we used the sequence-based method PSIPRED V3.0 [55] to estimate protein secondary structures. For the other HIV-1 proteins with available PDB data (Additional file 2: Table S5), we assessed protein secondary structures using both PSIPRED V3.0 [55] and 2Struc [56]. 2Struc is a software platform which integrates 8 PDB-based methods: DSSP_CONT, DSSP, KAKSI, PALSSE, P-SEA, STICKS, STRIDE and XTLSSTR [56]. Alpha-helix, beta-strand and random-coil structures were estimated using the majority voting of above 9 methods. Prediction similarities between these 9 methods are shown in Additional file 1: Figure S23.

Protein intrinsic disorder

Protein disordered regions are exploited by the virus to invade cellular host systems [30]; these regions are often structurally unstable without their partner molecules [57]. We estimated the intrinsically disordered structures of HIV-1 subtype B proteins using three software packages: MetaPrDOS [57], VSL2P [58] and PreDisorder v1.1 [59]. A disorder score (a numerical value between 0 and 1) of each amino acid position was estimated by 17 methods in these 3 software packages. An amino acid position was estimated as intrinsically disordered if its disorder score was above the cutoff value of 0.5 [57-59]. The intrinsically disordered positions were identified based on the majority voting of 17 methods. Prediction similarities between these 17 methods are shown in Additional file 1: Figure S24.

Solvent accessible surface area

We estimated protein solvent accessible surface areas (ASA) using Chimera V1.6.1 [60] (default parameters). Provided with PDB data in Additional file 2: Table S5, we calculated the ASA at each amino acid of all HIV-1 protein units. For each of the 20 amino acids, a distribution of its ASA scores over 15 HIV-1 proteins was obtained and the maximum ASA was identified therein. An amino acid at a specific position was considered buried if its ASA was lower than 25% of the maximum ASA for the corresponding amino acid [61] (Additional file 2: Table S8).

Phylogenetic analysis

Our phylogenetic analysis was performed using 1384 nucleotide genomic sequences of 14 HIV groups and pure subtypes (thus excluding CRFs), obtained from the earliest sampling time (one sequence per patient). To prepare the alignment, we also removed ambiguous regions containing multiple insertions, deletions and hypervariable positions (HXB2 index: 1126-1182, 6866-7003, 7106-7154, 7773-7842, 7981-8032, 8897-9383). Maximum-likelihood phylogenetic trees were obtained using the multi-threaded FastTree V2.1 [62]. Our software parameters were set to 100 bootstrap replicates, the fully optimized GTR (generalized time-reversible) model, the continuous gamma distribution and the exhaustive nearest-neighbor interchange approach. The consensus phylogenetic tree with bootstrap supports was obtained using the seqboot tool in Phylip V3.69 (http://evolution.genetics.washington.edu/phylip.html).

Ratio of non-synonymous and synonymous rate (dN/dS)

Following the protocol described in [26], we performed the dN/dS analysis for HIV-1 subtype B using 495 genomic sequences sampled from different treatment-naïve patients. We firstly constructed a maximum-likelihood phylogenetic tree using FastTree V2.1 [62] (parameters: continuous gamma distribution, generalized time-reversible (GTR) model). Provided with the constructed phylogenetic tree and the codon sequence dataset, we then applied HyPhy V2.1.0 [63] to estimate the non-synonymous (dN) and synonymous (dS) substitution rates. We employed the single likelihood ancestor counting (SLAC) model with the optimized GTR [63]. Ambiguous nucleotides were resolved by averaging over all possible states for the ancestral sequence reconstruction [63]. Statistical significance of dN/dS was measured by the continuous extension of binomial distributions [63]. The above procedure was also performed for HIV-1 subtype A1, C and CRF01_AE genomes using our sequence datasets (Additional file 2: Table S2).

Quantification of sequence diversity

Sequence diversity was calculated based on the pairwise nucleotide (NT) and amino acid (AA) comparisons [9,64]. When calculating the amino acid diversity of HIV genome, we concatenated the amino acid sequences of 15 HIV protein coding regions in the full-length genome. Suppose the sequence dataset D contains L sequences with N positions, genetic diversity at position n is calculated by: $GD(D_n) = 1 - \frac{2}{L(L-1)} \sum_{i=1}^{L} \sum_{j=i+1}^{L} \delta\left(D_{n,i} = D_{n,j}\right)$, where $D_{n,i}$ is the NT or AA form of the position n at the i$^{\text{th}}$ sequence in the dataset D, δ represents the Kronecker symbol, $\delta(D_{n,i} = D_{n,j})$ equals 1 if $D_{n,i}$ is identical to $D_{n,j}$; otherwise 0. Given the sequence dataset D, intra-clade genetic diversity *AGD(D)* is defined as the average genetic diversity of all positions:

$$AGD(D) = 1 - \frac{1}{N} \sum_{n=1}^{N} \frac{2}{L(L-1)} \sum_{i=1}^{L} \sum_{j=i+1}^{L} \delta\left(D_{n,i} = D_{n,j}\right).$$

Suppose two sequence datasets D1 and D2 aligned with the same reference genome have the number of sequences L_1 and L_2 respectively. The inter-clade genetic diversity between D1 and D2 is defined as:

$$RGD(D1, D2) = 1 - \frac{1}{N} \sum_{n=1}^{N} \frac{1}{L_1 \times L_2} \sum_{i=1}^{L_1} \sum_{j=1}^{L_2} \delta\left(D1_{n,i} = D2_{n,j}\right).$$

Furthermore, only positions for which less than 20% of sequences had gaps were considered and gaps were treated as missing data. Intra- and inter-clade genetic diversity was measured using one genomic sequence per patient, while intra-patient diversity was calculated using more than one genomic sequence sampled from individual patients. The Mann–Whitney *U* test was performed to compare the distributions of genetic diversity and a significant difference was identified if a p-value was lower than 0.05 [65]. Our Matlab implementation of genomic diversity analysis is available in Additional file 3.

Additional files

Additional file 1: Figures and tables. Figure S1. Gene maps and protein structures of HIV-1 and HIV-2. **Figure S2.** Distribution plots of nucleotide and AA diversity among HIV types, groups and subtypes. **Figure S3.** Distribution plots of AA diversity between HIV-1 subtype B/C and the other HIV groups/subtypes. **Figure S4.** Global distribution of HIV-1 genomic diversity. **Figure S5.** AA diversity along the full-length HIV genome. **Figure S6.** Global distribution of HIV-1 genomic diversity. **Figure S7.** Average AA diversity of HIV-1 protein clusters and number of HIV-human protein interactions. **Figure S8.** AA composition of HIV-1 subtype B genome, HIV-1 peptide-derived regions and sequences of HIV-derived peptide inhibitors. **Figure S9.** Average AA diversity of peptide-derived regions in HIV-1 subtype B. **Figure S10.** Solvent accessible surface area of peptide-derived regions in the HIV-1 subtype B genome. **Figure S11.** Protein intrinsic disorder scores of peptide-derived regions in the HIV-1 subtype B genome. **Figure S12.** Protein structure of the HIV-1 GP120-CD4-Fab 48d complex (PDB: 2B4C, 3U4E) and mapped GP120 peptide-derived inhibitors. **Figure S13.** GP41 structure and GP41-derived peptide inhibitors. **Figure S14.** HIV-1 Integrase tetramer and Integrase-derived peptide inhibitors. **Figure S15.** HIV-1 RT structure and RT-derived peptide inhibitors. **Figure S16.** HIV-1 Protease homodimer structure and protease-derived peptide inhibitors. **Figure S17.** HIV-1 Tat structure and Tat-derived peptide inhibitors. **Figure S18.** HIV-1 Vpr structure and Vpr-derived peptides. **Figure S19.** HIV-1 Rev tetramer structure and Rev-derived peptide inhibitors. **Figure S20.** Structure of HIV-1 Capsid monomer and Capsid-derived peptide inhibitors. **Figure S21.** HIV-1 Vif structure and Vif-derived peptide inhibitors. **Figure S22.** Distribution plots of AA diversity between the consensus and the circulating genomes, within circulating genomes. **Figure S23.** Prediction similarities of the consensus and the 9 protein secondary structure prediction methods. **Figure S24.** Prediction similarities of the consensus and 17 methods for protein intrinsically disorder prediction.

Additional file 2: Table S1. Average amino acid diversity of HIV monomeric and multimeric proteins. **Table S2.** Summary of average AA diversity, average dN, average dS and average dN/dS in the HIV-1 subtype A1, B, C and CRF 01_AE genomes. **Table S3.** Statistical of dN/dS, dN and dS distributions in the monomeric and multimeric protein groups of the HIV-1 subtype A1, B, C and CRF01_AE genomes. **Table S4.** Summary of 121 peptide inhibitors derived from HIV-1 proteins. **Table S5.** Summary of protein structures and PDB data for HIV-1 and HIV-2 proteins. **Table S6.** Percentages of deletions and insertions in the HIV-1 and HIV-2 full-length genomic sequence alignments. **Table S7.** Summary of antibody, CD4+, CD8+ T cell epitope positions. **Table S8.** Cutoffs for determining solvent exposed residues.

Additional file 3: Software. Our Matlab toolbox developed for data visualization, genomic diversity analysis and HIV genomic alignment. Our sequence datasets are also included.

Abbreviations

AA: Amino acid; ASA: Solvent accessible surface area; CA: Capsid; CRF: Circulating recombinant form; dN/dS: Ratio of non-synonymous to synonymous; EC_{50}: 50% effective concentration; GTR: Generalized time reversible model; LTR: Long terminal region; IC_{50}: 50% inhibitory concentration; IN: Integrase; MA: Matrix; MSA: Multiple sequence alignment;

NC: Nucleocapsid; NT: Nucleotide; OR: Odds ratio; ORF: Open reading frame; PDB: Protein data bank; PR: Protease; RT: Reverse transcriptase.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

GL, A-MV and KT designed the study and drafted the manuscript; GL, PL and NRF performed the phylogenetic analysis; GL and SP reviewed the peptide inhibitor data; GL and AV collected and analyzed the PDB data; GL, AP and RK collected and analyzed the immunological data; GL designed the software. All authors read and approved the final manuscript.

Acknowledgements

We thank Fossie Ferreira, Jasper Edgar Neggers, Soraya Maria Menezes and Tim Dierckx for technical assistance and valuable contributions to our analysis. This work was supported by the National Nature Science Foundation of China [81130015]; the National Basic Research Program of China [2014CB910500]; the Fonds voor Wetenschappelijk Onderzoek – Flanders (FWO) [PDO/11 to K.T., G069214N]; the European Community's Seventh Framework Programme (FP7/2007-2013) under the project "Collaborative HIV and Anti-HIV Drug Resistance Network (CHAIN)" [223131].

Author details

[1]Metabolic Syndrome Research Center, the Second Xiangya Hospital, Central South University, Changsha, Hunan, China. [2]Rega Institute for Medical Research, Department of Microbiology and Immunology, KU Leuven, Leuven, Belgium. [3]Department of Zoology, University of Oxford, Oxford OX1-3PS, UK. [4]Zhang IRU, RIKEN Institute Laboratories, Hirosawa 2-1, Wako-shi, Saitama, Japan. [5]Clinical and Molecular Infectious Disease Group, Faculty of Sciences and Mathematics, Universidad del Rosario, Bogotá, Colombia. [6]LIMI-LIP, Centro de Pesquisa Gonçalo Moniz, FIOCRUZ, Salvador-Bahia, Brasil. [7]Centro de Malária e Outras Doenças Tropicais and Unidade de Microbiologia, Instituto de Higiene e Medicina Tropical, Universidade Nova de Lisboa, Lisbon, Portugal.

Received: 11 July 2014 Accepted: 28 January 2015
Published online: 15 February 2015

References

1. Hemelaar J. The origin and diversity of the HIV-1 pandemic. Trends Mol Med. 2012;18:182–92.
2. Gaschen B, Taylor J, Yusim K, Foley B, Gao F, Lang D, et al. Diversity considerations in HIV-1 vaccine selection. Science. 2002;296:2354–60.
3. Frankel AD, Young JA. HIV-1: fifteen proteins and an RNA. Annu Rev Biochem. 1998;67:1–25.
4. Engelman A, Cherepanov P. The structural biology of HIV-1: mechanistic and therapeutic insights. Nat Rev Microbiol. 2012;10:279–90.
5. Acosta EG, Kumar A, Bartenschlager R. Revisiting dengue virus-host cell interaction: new insights into molecular and cellular virology. Adv Virus Res. 2014;88:1–109.
6. Perry CM. Elvitegravir/cobicistat/emtricitabine/tenofovir disoproxil fumarate single-tablet regimen (Stribild(R)): a review of its use in the management of HIV-1 infection in adults. Drugs. 2014;74:75–97.
7. Tilton JC, Doms RW. Entry inhibitors in the treatment of HIV-1 infection. Antiviral Res. 2010;85:91–100.
8. Fauci AS, Folkers GK, Dieffenbach CW. HIV-AIDS: much accomplished, much to do. Nat Immunol. 2013;14:1104–7.
9. Li G, Verheyen J, Rhee SY, Voet A, Vandamme AM, Theys K. Functional conservation of HIV-1 gag: implications for rational drug design. Retrovirology. 2013;10:126.
10. Stephenson KE, Barouch DH. A global approach to HIV-1 vaccine development. Immunol Rev. 2013;254:295–304.
11. Rolland M, Tovanabutra S, deCamp AC, Frahm N, Gilbert PB, Sanders-Buell E, et al. Genetic impact of vaccination on breakthrough HIV-1 sequences from the STEP trial. Nat Med. 2011;17:366–71.
12. Rolland M, Edlefsen PT, Larsen BB, Tovanabutra S, Sanders-Buell E, Hertz T, et al. Increased HIV-1 vaccine efficacy against viruses with genetic signatures in Env V2. Nature. 2012;490:417–20.
13. Arien KK, Vanham G, Arts EJ. Is HIV-1 evolving to a less virulent form in humans? Nat Rev Microbiol. 2007;5:141–51.
14. Herbeck JT, Rolland M, Liu Y, McLaughlin S, McNevin J, Zhao H, et al. Demographic processes affect HIV-1 evolution in primary infection before the onset of selective processes. J Virol. 2011;85:7523–34.
15. Rousseau CM, Birditt BA, McKay AR, Stoddard JN, Lee TC, McLaughlin S, et al. Large-scale amplification, cloning and sequencing of near full-length HIV-1 subtype C genomes. J Virol Methods. 2006;136:118–25.
16. Wang YE, Li B, Carlson JM, Streeck H, Gladden AD, Goodman R, et al. Protective HLA class I alleles that restrict acute-phase CD8+ T-cell responses are associated with viral escape mutations located in highly conserved regions of human immunodeficiency virus type 1. J Virol. 2009;83:1845–55.
17. Brown BK, Darden JM, Tovanabutra S, Oblander T, Frost J, Sanders-Buell E, et al. Biologic and genetic characterization of a panel of 60 human immunodeficiency virus type 1 isolates, representing clades A, B, C, D, CRF01_AE, and CRF02_AG, for the development and assessment of candidate vaccines. J Virol. 2005;79:6089–101.
18. Fernandez-Garcia A, Cuevas MT, Munoz-Nieto M, Ocampo A, Pinilla M, Garcia V, et al. Development of a panel of well-characterized human immunodeficiency virus type 1 isolates from newly diagnosed patients including acute and recent infections. AIDS Res Hum Retroviruses. 2009;25:93–102.
19. Kousiappa I, Van De Vijver DA, Kostrikis LG. Near full-length genetic analysis of HIV sequences derived from Cyprus: evidence of a highly polyphyletic and evolving infection. AIDS Res Hum Retroviruses. 2009;25:727–40.
20. Sanabani SS, Pessoa R, Soares de Oliveira AC, Martinez VP, Giret MT, de Menezes Succi RC, et al. Variability of HIV-1 genomes among children and adolescents from Sao Paulo, Brazil. PLoS One. 2013;8:e62552.
21. Fernandez-Garcia A, Revilla A, Vazquez-de Parga E, Vinogradova A, Rakhmanova A, Karamov E, et al. The analysis of near full-length genome sequences of HIV type 1 subtype A viruses from Russia supports the monophyly of major intrasubtype clusters. AIDS Res Hum Retroviruses. 2012;28:1340–3.
22. Henn MR, Boutwell CL, Charlebois P, Lennon NJ, Power KA, Macalalad AR, et al. Whole genome deep sequencing of HIV-1 reveals the impact of early minor variants upon immune recognition during acute infection. PLoS Pathog. 2012;8:e1002529.
23. Guyader M, Emerman M, Sonigo P, Clavel F, Montagnier L, Alizon M. Genome organization and transactivation of the human immunodeficiency virus type 2. Nature. 1987;326:662–9.
24. Korber B, Gaschen B, Yusim K, Thakallapally R, Kesmir C, Detours V. Evolutionary and immunological implications of contemporary HIV-1 variation. Br Med Bull. 2001;58:19–42.
25. van der Kuyl AC, Berkhout B. The biased nucleotide composition of the HIV genome: a constant factor in a highly variable virus. Retrovirology. 2012;9:92.
26. Snoeck J, Fellay J, Bartha I, Douek DC, Telenti A. Mapping of positive selection sites in the HIV-1 genome in the context of RNA and protein structural constraints. Retrovirology. 2011;8:87.
27. Mayrose I, Stern A, Burdelova EO, Sabo Y, Laham-Karam N, Zamostiano R, et al. Synonymous site conservation in the HIV-1 genome. BMC Evol Biol. 2013;13:164.
28. Santoro MM, Perno CF. HIV-1 Genetic Variability and Clinical Implications. ISRN Microbiol. 2013;2013:481314.
29. Hemelaar J, Gouws E, Ghys PD, Osmanov S. Isolation W-UNfH, Characterisation: Global trends in molecular epidemiology of HIV-1 during 2000-2007. AIDS. 2011;25:679–89.
30. Xue B, Mizianty MJ, Kurgan L, Uversky VN. Protein intrinsic disorder as a flexible armor and a weapon of HIV-1. Cell Mol Life Sci. 2012;69:1211–59.
31. Haynes BF, Gilbert PB, McElrath MJ, Zolla-Pazner S, Tomaras GD, Alam SM, et al. Immune-correlates analysis of an HIV-1 vaccine efficacy trial. Engl J Med. 2012;366:1275–86.
32. Boons E, Li G, Vanstreels E, Vercruysse T, Pannecouque C, Vandamme A-M, et al. A stably expressed llama single-domain intrabody targeting Rev displays broad-spectrum anti-HIV activity. Antiviral Research. 2014;112:91–102.
33. Rambaut A, Posada D, Crandall KA, Holmes EC. The causes and consequences of HIV evolution. Nat Rev Genet. 2004;5:52–61.
34. Falkowski PG, Fenchel T, Delong EF. The microbial engines that drive Earth's biogeochemical cycles. Science. 2008;320:1034–9.
35. Dunwell JM, Culham A, Carter CE, Sosa-Aguirre CR, Goodenough PW. Evolution of functional diversity in the cupin superfamily. Trends Biochem Sci. 2001;26:740–6.
36. Buck PM, Kumar S, Singh SK. On the role of aggregation prone regions in protein evolution, stability, and enzymatic catalysis: insights from diverse analyses. PLoS Comput Biol. 2013;9:e1003291.

37. Li G, Theys K, Verheyen J, Pineda-Pena A, Khouri R, Piampongsant S, et al. A new ensemble coevolution system for detecting HIV-1 protein coevolution. Biol Direct. 2015;10:1.
38. Zolla-Pazner S, Cardozo T. Structure-function relationships of HIV-1 envelope sequence-variable regions refocus vaccine design. Nat Rev Immunol. 2010;10:527–35.
39. Jager S, Cimermancic P, Gulbahce N, Johnson JR, McGovern KE, Clarke SC, et al. Global landscape of HIV-human protein complexes. Nature. 2012;481:365–70.
40. Burnett JC, Zaia JA, Rossi JJ. Creating genetic resistance to HIV. Curr Opin Immunol. 2012;24:625–32.
41. He Y. Synthesized peptide inhibitors of HIV-1 gp41-dependent membrane fusion. Curr Pharm Des. 2013;19:1800–9.
42. Roberts KE, Cushing PR, Boisguerin P, Madden DR, Donald BR. Computational design of a PDZ domain peptide inhibitor that rescues CFTR activity. PLoS Comput Biol. 2012;8:e1002477.
43. Durrant JD, McCammon JA. Molecular dynamics simulations and drug discovery. BMC Biol. 2011;9:71.
44. Liu S, Lu H, Niu J, Xu Y, Wu S, Jiang S. Different from the HIV fusion inhibitor C34, the anti-HIV drug Fuzeon (T-20) inhibits HIV-1 entry by targeting multiple sites in gp41 and gp120. J Biol Chem. 2005;280:11259–73.
45. Rolland M, Nickle DC, Mullins JI. HIV-1 group M conserved elements vaccine. PLoS Pathog. 2007;3:e157.
46. Kwong PD, Mascola JR, Nabel GJ. Broadly neutralizing antibodies and the search for an HIV-1 vaccine: the end of the beginning. Nat Rev Immunol. 2013;13:693–701.
47. Li G, Verheyen J, Theys K, Piampongsant S, Van Laethem K, Vandamme AM. HIV-1 Gag C-terminal amino acid substitutions emerging under selective pressure of protease inhibitors in patient populations infected with different HIV-1 subtypes. Retrovirology. 2014;11:79.
48. Pineda-Pena AC, Faria NR, Imbrechts S, Libin P, Abecasis AB, Deforche K, et al. Automated subtyping of HIV-1 genetic sequences for clinical and surveillance purposes: Performance evaluation of the new REGA version 3 and seven other tools. Infect Genet Evol. 2013;19:337–48.
49. Struck D, Lawyer G, Ternes AM, Schmit JC, Bercoff DP. COMET: adaptive context-based modeling for ultrafast HIV-1 subtype identification. Nucleic Acids Res. 2015;42:e144.
50. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32:1792–7.
51. Gouy M, Guindon S, Gascuel O. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol Biol Evol. 2010;27:221–4.
52. Hooft RW, Vriend G, Sander C, Abola EE. Errors in protein structures. Nature. 1996;381:272.
53. Pinney JW, Dickerson JE, Fu W, Sanders-Beer BE, Ptak RG, Robertson DL. HIV-host interactions: a map of viral perturbation of the host system. AIDS. 2009;23:549–54.
54. Llano A, Frahm N, Brander C. How to optimally define optimal cytotoxic T lymphocyte epitopes in HIV infection. HIV Mol Immunol. 2009;2009:3–24.
55. Buchan DW, Ward SM, Lobley AE, Nugent TC, Bryson K, Jones DT. Protein annotation and modelling servers at University College London. Nucleic Acids Res. 2010;38:W563–8.
56. Klose DP, Wallace BA, Janes RW. 2Struc: the secondary structure server. Bioinformatics. 2010;26:2624–5.
57. Ishida T, Kinoshita K. Prediction of disordered regions in proteins based on the meta approach. Bioinformatics. 2008;24:1344–8.
58. Peng K, Radivojac P, Vucetic S, Dunker AK, Obradovic Z. Length-dependent prediction of protein intrinsic disorder. BMC Bioinformatics. 2006;7:208.
59. Deng X, Eickholt J, Cheng J. PreDisorder: ab initio sequence-based prediction of protein disordered regions. BMC Bioinformatics. 2009;10:436.
60. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF Chimera–a visualization system for exploratory research and analysis. J Comput Chem. 2004;25:1605–12.
61. Levy ED. A simple definition of structural regions in proteins and its use in analyzing interface evolution. J Mol Biol. 2010;403:660–70.
62. Price MN, Dehal PS, Arkin AP. FastTree 2–approximately maximum-likelihood trees for large alignments. PLoS One. 2010;5:e9490.
63. Pond SL, Frost SD, Muse SV. HyPhy: hypothesis testing using phylogenies. Bioinformatics. 2005;21:676–9.
64. Spira S, Wainberg MA, Loemba H, Turner D, Brenner BG. Impact of clade diversity on HIV-1 virulence, antiretroviral drug sensitivity and drug resistance. J Antimicrob Chemother. 2003;51:229–40.
65. Li G: HIV genome-wide diversity, interaction and coevolution. Doctoral thesis. University of Leuven, Faculty of medicine; 2014. (https://lirias.kuleuven.be/handle/123456789/460408)

Submit your next manuscript to BioMed Central and take full advantage of:

- **Convenient online submission**
- **Thorough peer review**
- **No space constraints or color figure charges**
- **Immediate publication on acceptance**
- **Inclusion in PubMed, CAS, Scopus and Google Scholar**
- **Research which is freely available for redistribution**

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central

Kirchner *et al. Genome Medicine* (2016) 8:92
DOI 10.1186/s13073-016-0343-7

Genome Medicine

REVIEW Open Access

Recent advances in malaria genomics and epigenomics

Sebastian Kirchner, B. Joanne Power and Andrew P. Waters*

Abstract

Malaria continues to impose a significant disease burden on low- and middle-income countries in the tropics. However, revolutionary progress over the last 3 years in nucleic acid sequencing, reverse genetics, and post-genome analyses has generated step changes in our understanding of malaria parasite (*Plasmodium* spp.) biology and its interactions with its host and vector. Driven by the availability of vast amounts of genome sequence data from *Plasmodium* species strains, relevant human populations of different ethnicities, and mosquito vectors, researchers can consider any biological component of the malarial process in isolation or in the interactive setting that is infection. In particular, considerable progress has been made in the area of population genomics, with *Plasmodium falciparum* serving as a highly relevant model. Such studies have demonstrated that genome evolution under strong selective pressure can be detected. These data, combined with reverse genetics, have enabled the identification of the region of the *P. falciparum* genome that is under selective pressure and the confirmation of the functionality of the mutations in the *kelch13* gene that accompany resistance to the major frontline antimalarial, artemisinin. Furthermore, the central role of epigenetic regulation of gene expression and antigenic variation and developmental fate in *P. falciparum* is becoming ever clearer. This review summarizes recent exciting discoveries that genome technologies have enabled in malaria research and highlights some of their applications to healthcare. The knowledge gained will help to develop surveillance approaches for the emergence or spread of drug resistance and to identify new targets for the development of antimalarial drugs and perhaps vaccines.

Background

Malaria, caused by unicellular protozoan *Plasmodium* spp. parasites, is an ancient disease and remains a major threat to human health and wellbeing. Five species of *Plasmodium* are currently recognized as causing human malaria, of which the most lethal is *P. falciparum* (Pf). In 2015, the World Health Organization estimated that the maximal annual burden imposed by malaria, whilst decreasing, still stands at 214 million (range 149–303 million) cases that result in 438,000 (range 236,000–623,000) deaths [1]. Drug resistance to frontline antimalarials continues to arise and spread, exacerbated by slow progress in the introduction of alternatives. Properly efficacious vaccines remain a hope, not a likelihood. Against this background, genome-based research on malaria seeks to provide new avenues for therapeutic or prophylactic development based upon biological insights such as the identification of new drug targets and vaccine candidates.

The landmark of the completion of the genome sequence of a laboratory strain of Pf was achieved over a decade ago [2] (Fig. 1). This has since been accompanied, thanks to plummeting costs and advances in next-generation sequencing (NGS) technologies, by the whole-genome sequencing (WGS) of a wide range of species representing all the major clades of the genus, although the genomes of all known human infectious *Plasmodium* species remain to be sequenced [3]. However, the combination of NGS and WGS has enabled the development of innovative large-scale genomic studies, for example, for genomic epidemiology [4]. Such population genomics, fueled by collaborative consortia (for example, the Malaria Genomic Epidemiology Network (MalariaGEN; http://www.malariagen.net), have allowed the dynamics of global and local population

---

* Correspondence: AndyWaters@Glasgow.ac.uk
Wellcome Trust Centre for Molecular Parasitology, Institute of Infection, Immunity and Inflammation, College of Medical Veterinary & Life Sciences, University of Glasgow, Glasgow G12 8TA, UK

© 2016 The Author(s). **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

structures to be assessed and adaptive change in parasite genomes to be monitored in response to threats, such as artemisinin (ART). This is especially true for single-nucleotide polymorphisms (SNPs), and while other aspects of genome variation (such as indels and copy number variation) might currently lag behind, the gaps in the database are known and are firmly in the sights of researchers.

The template *Plasmodium* genomes have provided the substrate for the application of an explosion of other post-genome survey technologies that have been largely exclusively applied to Pf, such as transcriptomics, proteomics, metabolomics, and lipidomics, and that map the general and stage-specific characteristics of malaria parasites. These data are warehoused in expensive but critical community web sites such as PlasmoDB (http://www.Plasmodb.org). This in turn has been exploited by ever improving forward and reverse genetic capabilities to assign function to genes, steadily reducing the >60 % of genes of unknown function that were originally catalogued [2]. Advances that will be highlighted in this review include: the unraveling of the molecular mechanisms of parasite resistance to ART; the functional identification of some of the histone-modifying enzymes that write the epigenetic code (such as Pf histone deacetylase 2 (PfHDA2)) and the proteins that read it (such as Pf heterochromatin protein 1 (PfHP1)) that, with others (such as RNaseII), play a significant role in the regulation of antigenic variation and commitment to sexual development.

Furthermore, the genomes of the host and of a growing number of mosquito vectors have been characterized in both increasing number and depth, permitting meta-analyses of these genomes in combination with *Plasmodium* infection. These studies have revealed important loci associated with resistance to the malaria parasite in the host and vector, respectively [5, 6], and indicate the genomic hotspots in the genetic arms race that malaria has stimulated.

We also review the recent advances in this very active area of malaria genomics and control of gene expression and emphasize any benefits that these advances may have for the development of therapies and interventions (Table 1).

Human genomics

The infrastructure required to effectively collect, collate, and analyze large genomes for epidemiological studies (that is, genome-wide association studies (GWAS)) is so costly that it is best achieved in consortia. These can work at such a scale that analyses are powered to a degree that GWAS findings become more certain and the global context of the effect of, for example, human genetics on susceptibility to malaria is more reliably resolved. The African Genome Variation Project recognizes the significant diversity of ethnicities and, therefore, genotypes and, through WGS, imputation, and SNP mapping, seeks to build a database through which disease incidence and outcome can be reliably associated with haplotypes [7]. Already, such wider analyses have confirmed SNP associations with five well-known traits, including hemoglobinopathies and glucose-6-phosphate dehydrogenase (G6PD) deficiency, but have refuted 22 others that had been linked by smaller-scale studies [8]. This study also showed opposing effects of G6PD on different fatal consequences of malaria infection, revealing a hitherto unsuspected complexity of associations. Ongoing analyses have revealed new, although unsurprising, examples of haplotypes of loci associated with protection from severe malaria, such as the glycophorin locus on human chromosome 4 [8, 9].

Vector genomics

In Africa, malaria is mainly transmitted by female *Anopheles gambiae* (Ag) mosquitoes. Approaches to understanding the role of Ag mosquito genomics in malaria transmission have been similar to those of the African Genome Variation Project. Thus, the Ag1000G project (https://www.malariagen.net/projects/ag1000g) involves 35 working groups that have sampled Ag mosquitos from 13 malaria endemic countries and that aim to establish the levels of Ag genome diversity, establish population structures, and link these to the ecology of disease transmission. The *Anopheles* vector genome is very dynamic. Comparative vector genomics has revealed rapid gene gain and loss compared with *Drosophila* and significant intragenus diversity and mixing in genes involved in both insecticide resistance and antimalarial immunity [10, 11]. The nature and extent of such diversity precludes the application of classic GWAS approaches and a novel approach of phenotype-driven, pooled sequencing coupled with linkage mapping in carefully selected founder colonies has been used to map vector phenotypes. This study recently revealed *TOLL11* as a gene that protects African mosquitoes against Pf infection [6].

Parasite genomics

Full genome sequences are now available for many strains of Pf [2], *Plasmodium vivax* [12], and *Plasmodium knowlesi* [13] among the human infectious parasites. Primate and rodent infectious species that are frequently used as model parasites have also been sequenced and include *Plasmodium berghei* (Pb), *Plasmodium cynomolgi, Plasmodium chabaudi* and *Plasmodium yoelii* [14]. Recently, the genomes of seven further primate infectious species have become available, demonstrating the close relationship between Pf and

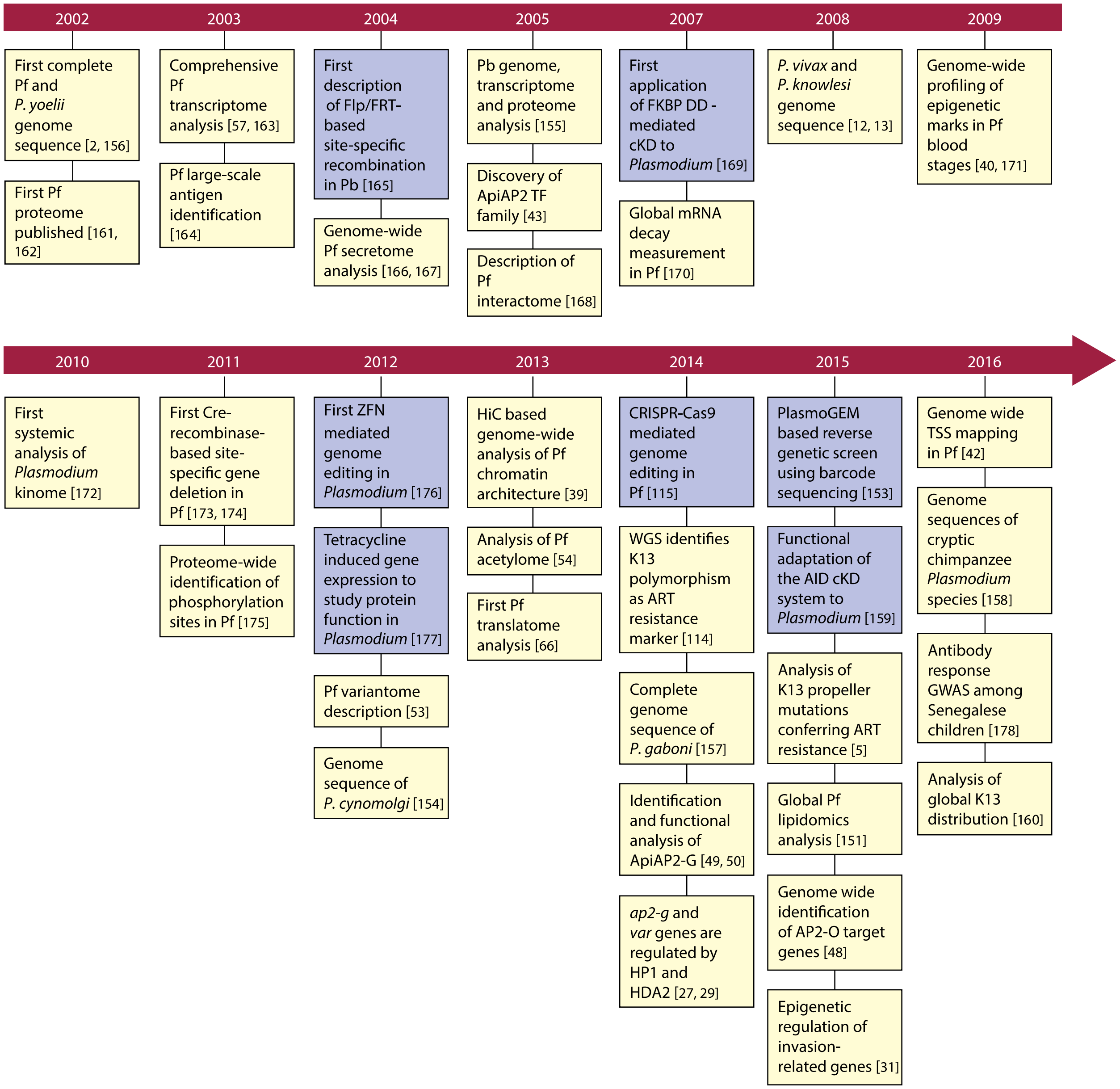

2002
First complete Pf and P. yoelii genome sequence [2, 156]
First Pf proteome published [161, 162]
2003
Comprehensive Pf transcriptome analysis [57, 163]
Pf large-scale antigen identification [164]
2004
First description of Flp/FRT-based site-specific recombination in Pb [165]
Genome-wide Pf secretome analysis [166, 167]
2005
Pb genome, transcriptome and proteome analysis [155]
Discovery of ApiAP2 TF family [43]
Description of Pf interactome [168]
2007
First application of FKBP DD - mediated cKD to Plasmodium [169]
Global mRNA decay measurement in Pf [170]
2008
P. vivax and P. knowlesi genome sequence [12, 13]
2009
Genome-wide profiling of epigenetic marks in Pf blood stages [40, 171]
2010
First systemic analysis of Plasmodium kinome [172]
2011
First Cre-recombinase-based site-specific gene deletion in Pf [173, 174]
Proteome-wide identification of phosphorylation sites in Pf [175]
2012
First ZFN mediated genome editing in Plasmodium [176]
Tetracycline induced gene expression to study protein function in Plasmodium [177]
Pf variantome description [53]
Genome sequence of P. cynomolgi [154]
2013
HiC based genome-wide analysis of Pf chromatin architecture [39]
Analysis of Pf acetylome [54]
First Pf translatome analysis [66]
2014
CRISPR-Cas9 mediated genome editing in Pf [115]
WGS identifies K13 polymorphism as ART resistance marker [114]
Complete genome sequence of P. gaboni [157]
Identification and functional analysis of ApiAP2-G [49, 50]
ap2-g and var genes are regulated by HP1 and HDA2 [27, 29]
2015
PlasmoGEM based reverse genetic screen using barcode sequencing [153]
Functional adaptation of the AID cKD system to Plasmodium [159]
Analysis of K13 propeller mutations conferring ART resistance [5]
Global Pf lipidomics analysis [151]
Genome wide identification of AP2-O target genes [48]
Epigenetic regulation of invasion-related genes [31]
2016
Genome wide TSS mapping in Pf [42]
Genome sequences of cryptic chimpanzee Plasmodium species [158]
Antibody response GWAS among Senegalese children [178]
Analysis of global K13 distribution [160]

Fig. 1 Major advances in omics-related fields. This figure highlights landmark studies providing key insights into parasite makeup, development, and pathogenesis (*yellow boxes*) as well as crucial technical advances (*blue boxes*) since the first *Plasmodium* genomes were published in 2002 [2, 5, 12, 13, 27, 29, 31, 39, 40, 42, 43, 48–50, 53, 54, 57, 66, 114, 115, 151, 153–178]. *AID* auxin-inducible degron, *ART* artemisinin, *cKD* conditional knockdown, *CRISPR* clustered regularly interspaced short palindromic repeats, *DD* destabilization domain, *K13* kelch13, *Pb P. berghei*, *Pf P. falciparum*, *TSS* transcription start site, *TF* transcription factor, *ZNF* zinc finger nuclease

chimpanzee infectious species [15]. The typical *Plasmodium* genome consists of 14 linear chromosomes of aggregate size of approximately 22 megabases encoding >5000 protein-coding genes. The core, conserved genome of about 4800 such genes occupies the central chromosomal regions whilst the multigene families (at least some of which are associated with antigenic variation) are largely distributed to the subtelomeric regions. Non-coding RNA (ncRNA) genes [16] and antisense transcription [17, 18] are being catalogued in Pf but this catalogue probably remains incomplete as only blood-stage parasites have been seriously investigated in this regard and ncRNAs remain largely of unknown significance.

One key feature of Pf is its evolution in the face of human-imposed selection pressures in the form of drugs and potentially vaccines. Such pressure has consistently resulted in the emergence of drug-resistant parasites.

Table 1 Key advances from recent omics studies

Research area	Key findings	Reference(s)
Genome sequencing	WGS and refinement of genome sequences of primate and rodent infectious *Plasmodium* species	[13, 154–157, 179]
	Genome sequence of cryptic chimpanzee *Plasmodium* species	[158]
Gene expression regulation	Discovery of AP2-G as master regulator of gametocytogenesis in Pf and Pb	[49, 50]
	AP2-G2 functions as a transcription regulator driving gametocyte development	[51]
	Functional analysis of Pf lifecycle dynamics using transcriptional profiling	[180]
	Genome-wide TSS mapping in Pf	[42]
	Acetylome analysis reveals post-translational PTM of ApiAP2 transcription factors	[55]
	Genome-wide analysis of AP2-O target genes	[48]
Post-transcriptional/translational control	Genome-wide analysis reveals features of the Pf translatome	[59, 66]
	Analysis of the Pb maternal repressome	[64]
Epigenetic control	SETvs functions as a regulator of *var* gene silencing in Pf	[98]
	Involvement of HP1 in clonal variant gene expression and sexual development	[29]
	HDA2 mediates *var* gene silencing and sexual commitment in Pf	[27]
	BDP1 regulates expression of invasion-related genes	[31]
	Profiling of histone PTMs in Pf	[22]
Gene silencing/editing	First description of the use of ZFN and CRISPR-Cas9 genome editing in *Plasmodium*	[115, 116, 140]
	Functional adaptation of an AID-based conditional knockdown system in Pb	[159]
Antibiotic resistance	Analysis of KELCH13-propellermutations in *Plasmodium* field isolates confirms relation to artemisinin resistance	[115, 116]
	Identification of UPR pathway involvement in *Plasmodium* artemisinin resistance	[123, 126]
Genome structure	First profiling of genome-wide chromosome interactions	[39]
	Strong correlation between spatial genome structure and gene expression	[37]
ncRNAs	PfRNaseII exonuclease is involved in group A *var* gene silencing	[84]
	Genome-wide search revealed over 1000 novel ncRNAs in Pf	[181]
	Strand-specific RNA sequencing reveals widespread transcription of antisense RNAs in Pf	[16–18]
	Antisense lncRNAs regulate expression of cognate *var* genes in Pf	[87]
Virulence/transmission	Vector transmission of *P. chabaudi chabaudi* impacts on parasite virulence	[182]
	Transcriptome changes during early phase of mosquito infection	[183]
Development	Global phosphoproteome analysis identified PKG as major intracellular Ca^{2+} level regulator	[184]
	PKG regulates *Plasmodium* egress and evasion	[185]
	Systematic subtractive bioinformatic analysis identified sex-specific stage V gametocyte protein marker	[186]
	Protein phosphatases are key regulators of *Plasmodium* development	[187]
Population genomics	Identification of a novel malaria resistance locus through GWAS in African children	[5]
	Genomic epidemiology study revels global KELCH13 mutation distribution	[160]
Metablomics	Lipid metabolism analysis in Pf	[151]

AID auxin-inducible degron, *CRISPR* clustered regularly interspaced short palindromic repeats, *GWAS* genome-wide association study, *lncRNA* long non-coding RNA, *ncRNA* non-coding RNA, *Pb P. berghei*, *Pf P. falciparum*, *PTM* post-translational modification, *SNP* single-nucleotide polymorphism, *TSS* transcription start site, *UPR* unfolded protein response, *WGS* whole-genome sequencing, *ZFN* zinc finger nuclease

There is a huge potential global reservoir of genome variation upon which selection may act. In an initial analysis of 227 parasite samples collected at six different locations in Africa, Asia, and Oceania, MalariaGEN, the Oxford-based genomic epidemiology network, identified more than 86,000 exonic SNPs. This initial SNP catalogue is described in detail by Manske and colleagues [19]. Currently (27 July 2016), the MalariaGEN database states that for the Pf Community Project, it has data on 3488 samples from 43 separate locations in 23 countries and the number of high-quality, filtered exonic SNPs has increased to more than

900,000. All of this variation is diversity, which in turn can be selected for fitter and perhaps more deadly parasites. Modern NGS and WGS have enabled comparative and population genomics approaches that have been used to reveal important features of emerging parasite populations, for example, in response to drugs.

Parasite development and pathogenesis

Within their mammalian host and mosquito vector *Plasmodium* parasites complete a remarkable lifecycle, alternating between asexual and sexual replication (Fig. 2). Throughout the *Plasmodium* lifecycle, regulation of gene expression is orchestrated by a variety of mechanisms, including epigenetic, transcriptional, post-transcriptional, and translational control of gene expression. Owing to the absence of most canonical eukaryotic transcription factors in the *Plasmodium* genome [2], epigenetic control has long been recognized to play an important role in gene expression regulation.

Epigenetics lies at the very heart of gene expression, regulating access of the transcriptional machinery to chromatin [20] via (1) post-translational modifications (PTMs) of histones, (2) nucleosome occupancy, and (3) global chromatin architecture. In the past decade, various histone PTMs have been identified throughout the *Plasmodium* lifecycle (reviewed in [21]) and the existing catalog of modifications in Pf was recently extended to 232 distinct PTMs, 88 unique to *Plasmodium* [22]. The majority of detected PTMs show dynamic changes across the intra-erythrocytic development cycle (IDC), likely mirroring changes within chromatin organization linked to its transcriptional status. Methylation and acetylation of N-terminal histone tails are by far the most studied regulatory PTMs, linked either to a transcriptionally active chromatin structure (that is, euchromatin) or to transcriptionally inert heterochromatin. In Pf, various genes encoding putative epigenetic modulators (that is, proteins catalyzing either the addition or removal of histone PTM marks) have been identified [23], but only a few have been subjected to more detailed investigation [24, 25]. Many of the histone modifiers are essential for *Plasmodium* development, making them a promising target for antimalarial drugs [26]. In Pf, conditional knockdown of HDA2, a histone lysine deacetylase (HDAC) catalyzing the removal of acetyl groups from acetylated histone 3 lysine 9 (H3K9ac), resulted in elevated H3K9ac levels in previously defined heterochromatin regions [27]. H3K9ac is an epigenetic mark associated with transcriptionally active euchromatin [28] and HDA2 depletion resulted in the transcriptional activation of genes located in heterochromatin regions, leading to impaired asexual growth and an increased gametocyte conversion [27]. Interestingly, genes found to be dysregulated by HDA2 knockdown are also known to be associated with HP1, a key epigenetic player binding to tri-methylated H3K9 (H3K9me3), linked to transcriptionally repressed chromatin. Strikingly, conditional knockdown of PfHP1 recapitulated, to a much greater extent, the phenotype observed in HDA2-knockdown mutants [29]. HP1 is believed to act as a recruitment platform for histone lysine methyltransferases (HKMTs), required for maintenance and spreading of H3K9me3 marks [30], which is consistent with the reduction of H3K9me3 observed in HP1 knockdown cells [29]. In addition, bromodomain protein 1 (BDP1) was found to bind to H3K9ac and H3K14ac marks within transcription start sites (TSSs) in Pf, among them predominantly invasion-related genes (Fig. 3a), and BDP1-knockdown parasites consistently failed to invade new erythrocytes. BDP1 also appears to act as a recruitment platform for other effector proteins such as BDP2 and members of the apicomplexan AP2 (ApiAP2) transcription factor family [31].

In addition to histone PTMs, nucleosome organization plays a critical role in gene expression regulation in *Plasmodium*. In general, heterochromatin is substantially enriched in nucleosomes compared with euchromatin [32] and active promoters and intergenic regions in Pf show markedly reduced nucleosome occupancy [33]. In addition, common transcript features such as TSSs, transcription termination sites, and splice donor/acceptor sites show clearly distinguishable nucleosome positioning in Pf [34], but previously described dynamic changes in nucleosome positioning [32] appeared to be mostly restricted to TSSs during the IDC [34]. Uniquely in *Plasmodium* spp., canonical histones in intergenic regions are replaced by histone variant H2A.Z [28], which, in concert with the apicomplexan-specific H2B.Z, establishes a H2A.Z/H2B.Z double-variant nucleosome subtype enriched at AT-rich promoter regions and correlates with open chromatin and active gene transcription [35].

Within the confined space of the nucleus, chromosomes are tightly packed into a three-dimensional structure. This three-dimensional architecture allows interaction between otherwise distant chromatin regions possessing regulatory function and facilitates contacts with other nuclear sub-compartments such as the nucleolus and the nuclear envelope [36]. Until recently, knowledge of the chromosome architecture and chromatin interactions in *Plasmodium* was mostly restricted to single genomic loci based on fluorescence *in situ* hybridization experiments [37]. However, recent advances in deep-sequencing technologies [38] have for the first time enabled the genome-wide profiling of chromosome interactions at kilobase resolution in *Plasmodium* [37, 39].

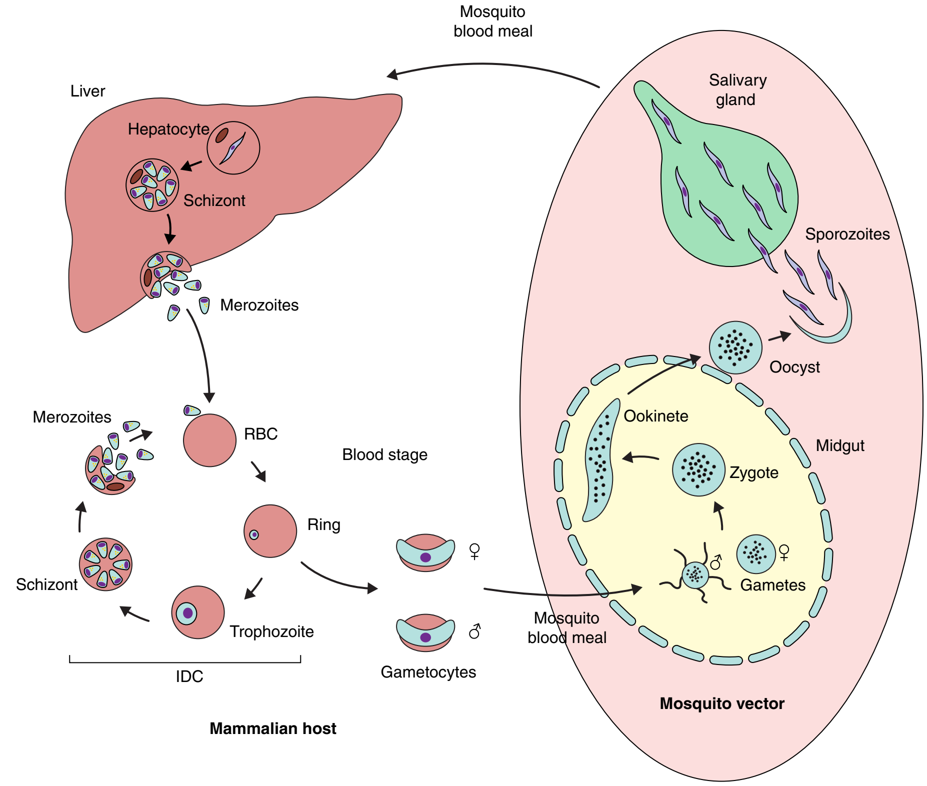


Fig. 2 *Plasmodium* life cycle. After a mosquito bite, malaria parasites are deposited into the host's skin and within minutes are carried via the bloodstream into the liver, where through asexual proliferation within the hepatocytes tens of thousands of merozoites are produced. Following hepatocyte rupture, merozoites are released into the bloodstream where they can invade the host's red blood cells (*RBC*), leading to the initiation of the intra-erythrocytic development cycle (*IDC*). During the IDC (lasting about 48–72 h in human and about 24 h in rodent malaria parasites), *Plasmodium* parasites multiply asexually through the completion of several morphologically distinct stages within the RBCs. After RBC invasion, malaria parasites develop via the ring and trophozoite stage into schizonts, each containing a species-specific number of merozoites (typically 10–30). Upon schizont rupture, merozoites are released into the bloodstream, where they can invade new RBCs and initiate a new IDC. However, a small fraction of ring-stage parasites sporadically differentiate into male or female gametocytes, which are responsible for initiating transmission back to the mosquito. Through another mosquito blood meal gametocytes are taken up into the mosquito midgut where they are activated and form male (eight per gametocyte) and female (one) gametes. Following fertilization, the zygote undergoes meiosis (and therefore true sexual recombination) and develops into a motile, tetraploid ookinete that traverses the midgut and forms an oocyst. Via another round of asexual proliferation inside the oocyst several thousands of new haploid sporozoites are generated that, upon their release, colonize the mosquito salivary glands, poised to initiate a new infection of another mammalian host

In contrast to other eukaryotic organisms, the Pf nucleus appears to lack clearly defined chromosome territories and chromatin interactions are mainly restricted to intra-chromosomal contacts showing a clear distance-related dependency [37, 39]. Inter-chromosomal interactions are mostly absent in Pf and restricted to centromeres, telomeres, ribosomal DNA (rDNA) loci, and internal as well as subtelomeric-localized *var* genes (further discussed in the next section). This observed clustering appears to coincide with transcriptional activity of each cluster. Interestingly, using three-dimensional chromatin modeling, the highly transcribed rDNA genes were proposed to be localized to the nuclear periphery, which was previously mainly associated with transcriptionally silenced heterochromatin [40], indicative of perinuclear transcriptionally active compartments [37].

Transcription itself is initiated through binding of the transcriptional machinery to promoter regions in the nucleus, resulting in the synthesis of pre-mRNA molecules, which, following extensive processing and nuclear export, leads to the accumulation of mature mRNAs in the parasite cytosol [41]. A recent study found evidence for stage-specific transcription initiation from distinct TSSs of otherwise identical transcriptional units, giving rise to developmentally regulated mRNA isoforms [42]. While most canonical eukaryotic transcription factors are absent from the *Plasmodium* genome [2], the ApiAP2 family of DNA-binding proteins comprises by far the largest group of transcription factors in malaria parasites [43]. A collection of ApiAP2 proteins is expressed throughout all stages of the IDC [44], while other ApiAP2 proteins are expressed outside the IDC

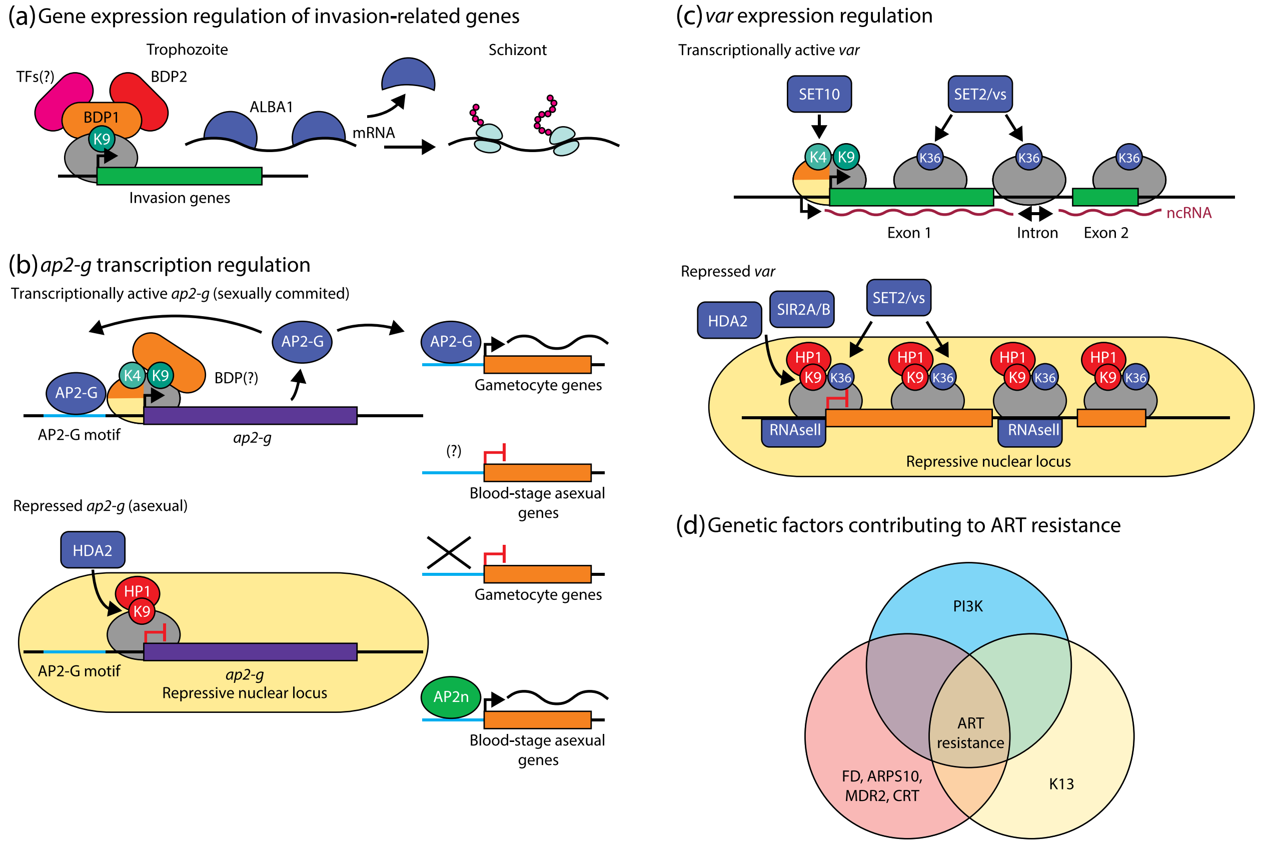

(a) Gene expression regulation of invasion-related genes
Trophozoite
Schizont
TFs(?)
BDP2
BDP1
K9
ALBA1
mRNA
Invasion genes
(b) ap2-g transcription regulation
Transcriptionally active ap2-g (sexually commited)
AP2-G
K4
K9
BDP(?)
AP2-G motif
ap2-g
Gametocyte genes
(?)
Blood-stage asexual genes
Repressed ap2-g (asexual)
HDA2
HP1
Repressive nuclear locus
AP2n
(c) var expression regulation
Transcriptionally active var
SET10
SET2/vs
K36
ncRNA
Exon 1
Intron
Exon 2
Repressed var
SIR2A/B
RNAseII
(d) Genetic factors contributing to ART resistance
PI3K
ART resistance
FD, ARPS10, MDR2, CRT
K13

Fig. 3 Malaria parasite genomic components involved in pathogenesis. **a** The expression of invasion-related genes is regulated through epigenetic and post-transcriptional mechanisms. Bromodomain protein 2 (*BDP2*) binds to H3K9ac marks within the promoter region of genes associated with red blood cell (RBC) invasion (as well as other gene families not depicted here [31]), enabling their transcription. This is likely achieved through the recruitment of BDP1 and transcription factors (*TFs*) of the ApiAP2 family. Following transcription during the trophozoite stage, mRNAs encoding invasion-related proteins are bound by ALBA1 functioning as translation repressor. After progression to the schizont stage, ALBA1 is released, allowing the timely synthesis of proteins required for merozoite invasion of RBCs. **b** Experimental findings either directly from studies on *ap2-g* or from epigenetically regulated *var* genes are suggestive of an epigenetically controlled mechanism regulating *ap2-g* transcription. In sexually committed parasites, *ap2-g* is characterized by H3K4me2/3 and H3K9ac histone marks and most likely contains histone variants H2A.Z and H2B.Z located in its promoter region. BDPs are believed to bind to H3K9ac, facilitating *ap2-g* transcription. ApiAP2-G drives expression of genes required for sexual development through binding to a 6/8-mer upstream DNA motif. *ap2-g* expression itself is believed to be multiplied through an autoregulatory feedback loop where ApiAP2-G binds to its own promoter that also contains ApiAP2-G motifs. In asexual blood-stage parasites, *ap2-g* is transcriptionally silenced by heterochromatin protein 1 (*HP1*) binding to H3K9me3 histone marks (located in repressive loci in the nuclear periphery). Histone deacetylase 2 (*HDA2*) catalyzes the removal of H3K9ac from active *ap2-g*, facilitating *ap2-g* silencing. **c** Monoallelic expression of one of the approximately 60 members of the erythrocyte membrane protein 1 (EMP1)-encoding *var* genes is regulated through epigenetic silencing of all but one *var* gene copy. The active *var* is marked by euchromatin post-translational modifications H3K4me2/3 and H3K9ac and histone variants H2A.Z/H2B.Z located in its promoter region, as well as H3K36me3 covering the whole *var* gene body but absent from the promoter region. Transcription of noncoding RNAs associated with the active *var* gene is facilitated by upstream as well as intronic promoters. All other silenced *var* genes cluster into perinuclear repressive loci and are characterized by HP1 binding to H3K9me3 marks. *var* gene silencing also involves SET2/vs-dependent placing of H3K36me3 histone marks in promoter regions and is marked by the absence of non-coding RNAs, likely safeguarded through RNaseII exonuclease activity. In addition, other histone code modulators such as HDA2, SET10, and SIR2A/B are likely involved in epigenetic *var* gene regulation. **d** Mutations in *kelch13* (K13) were found to be the major contributors to artemisinin (*ART*) resistance identified in drug-resistant parasites in the laboratory as well as in field isolates. *kelch13* mutations appear to arise in a complex array of background mutations (that is, mutations in genes encoding ferredoxin (*FD*), apicomplast ribosomal protein S10 (*ARPS10*), multidrug resistance protein 2 (*MDR2*), and chloroquine resistance transporter (*CRT*)), not yet detected in African parasites. In addition, elevated phosphatidylinositol-3-kinase (*PI3K*) levels have been observed in ART-resistant parasites and PI3K signaling has been implicated to impact on the unfolded protein response observed in ART-resistant parasites. H2A.Z/H2B.Z, *orange/yellow-paired quarter circles*; H3K4me2/3, *light green circles*; H3K9ac, *dark green circles*; H3K9me3, *red circles*; H3K36me3, *blue circles*; canonical nucleosomes, *grey globes*; ApiAP2-G binding motif; *light blue line*; ncRNAs, *wobbly red lines*; mRNAs, *wobbly black lines. AP2n* other TFs belonging to the ApiAP2 DNA binding protein family, *ncRNA* non-coding RNA, *TFs* transcription factors

[45–47]. ApiAP2s appear to be among the main drivers of developmental progress throughout most *Plasmodium* lifecycle stages and their disruption abolishes or greatly reduces parasite development [45, 46]. They bind in a sequence-specific fashion to motifs generally distributed upstream of open reading frames (ORFs) and individual AP2s may have widespread influence; PfAP2-O has been shown to bind upstream of >500 genes (roughly 10 % of the parasite ORFs), potentially influencing a wide range of cellular activities [48].

Through forward genetic screens and comparative genomics, ApiAP2-G was discovered to function as a conserved master regulator of sexual commitment in Pf and Pb. ApiAP2-G binds to a conserved 6/8-mer nucleotide motif enriched upstream of gametocyte-specific genes and *ap2-g* itself, leading to an autoregulatory feedback loop [49, 50] (Fig. 3b). ApiAP2-G2, another ApiAP2 family member, acts downstream of ApiAP2-G during sexual development, functioning as a transcriptional repressor blocking expression of genes required for asexual development and influencing gametocyte sex ratios [50, 51]. During the asexual IDC, *ap2-g* displays characteristics of epigenetically silenced heterochromatin, such as H3K9me3 marks, binding to HP1 and localization to the nuclear periphery (reviewed in [52]) (Fig. 3b). However, the previously mentioned knockdowns of both PfHDA2 and HP1 resulted in increased gametocyte conversion, likely as a direct consequence of the loss of H3K9me3 marks and H3K9 hyperacetylation leading to *ap2-g* transcriptional activation [27, 29]. This opens the possibility of a bet-hedging mechanism for sexual commitment in *Plasmodium*, regulating stochastic, low-level activation of *ap2-g* sensitive to environmental stimuli, as has been shown for several blood-stage-expressed genes [52, 53]. PTMs such as lysine acetylation are not restricted to histones and a recent study has demonstrated that the “acetylome” impacts >1000 proteins and intriguingly is highly enriched in the ApiAP2 transcription factor family [54, 55], although the functional consequences of these PTMs have yet to be established.

Following their synthesis, eukaryotic mRNAs are processed and are finally translated by the ribosomal machinery. Translation has long been a focus of malaria research, not only because it represents a promising target for antimalarial drugs but also for its potential regulatory features [56]. The lack of correlation between transcript and protein level observed throughout the *Plasmodium* lifecycle has fueled researchers’ interest in post-transcriptional and translational control for decades [57]. Many features of post-transcriptional/translational control in malaria parasites are similar to the mechanisms found in other eukaryotes [41]. However, the advent of ribosome profiling [58] has enabled in-depth genome-wide analysis of the *Plasmodium* translatome. Throughout the IDC, transcription and translation are tightly coupled and only 8 % (approximately 300 transcripts) of the transcriptome was found to be translationally regulated [59]. These genes were found to be involved in merozoite egress and invasion, and while transcript level peaked during the late stages of the IDC, maximal translation was observed during the early ring stage. This observation resembles a general feature of gene expression in *Plasmodium*, whereby for a set of genes transcription and translation are uncoupled and mRNA translation occurs during a later developmental time point when compared with maximal transcriptional activity, most notably in female gametocytes [46, 60–64]. This is especially true for genes required for developmental progression and provides the parasite with the capability of rapid and timely protein synthesis without the need for preceding de novo mRNA synthesis. Recently, PfALBA1, a member of the DNA/RNA-binding Alba protein family, was postulated to act as master regulator during the Pf IDC, controlling translation of invasion-related transcripts (Fig. 3a) as well as regulating mRNA homeostasis of approximately 100 transcripts in blood-stage parasites [65]. In contrast to findings by Caro and colleagues [59], an earlier study using polysome profiling found a discrepancy between steady-state mRNA level and polysome-associated mRNAs among 30 % of genes (1280 transcripts) during the Pf IDC, indicative of translationally controlled gene expression [66]. Additionally, the results of this study, as well the findings of others, suggest upstream ORF translation and stop codon read-through in Pf [67–69], but the genome-wide extent of such mechanisms in *Plasmodium* spp. remains controversial [59]. Hence, expansion of these studies to other parasite life stages, such as the gametocyte, where translational control is firmly established, would surely give further insights into the extent of translation regulation in *Plasmodium*.

In addition to canonical protein-coding mRNAs, a vast number of genes encoding different ncRNAs have been identified within the *Plasmodium* genome in recent years, which are believed to exert a variety of regulatory functions (reviewed in [70]). Circular RNAs (circRNAs) are among the newest members of the still expanding catalogue of existing ncRNAs in *Plasmodium* [17]. Host microRNAs (miRNAs) have been shown to regulate parasite translation [71], and circRNAs therefore might act as sponge for host miRNAs, a mechanism described in other organisms [72]. Recent studies have especially increased our knowledge of the role of ncRNAs in *var* gene regulation (discussed in the next section) but, nevertheless, the biological role of the vast majority of these ncRNA species remains unclear.

Immune evasion

In their attempts to occupy a diverse range of host environments, protozoan parasites of the *Plasmodium* genus have evolved a plethora of molecular mechanisms to evade the host adaptive immune response. The host immune response to *Plasmodium* infection is dependent upon both host and parasite genomics and the developmental stage and phenotype of the invading parasite [73–75]. In the best-studied example in *Plasmodium*, virulence of Pf is attributed largely to monoallelic expression of just one of approximately 60 *var* genes that encode variant copies of the surface antigen, Pf erythrocyte membrane protein 1 (PfEMP1). The ability to switch expression from one *var* gene to another enables the invading parasite to alternate between phenotypes of variable cytoadherent and immunogenic properties [76–78]. PfEMP1 proteins are expressed at parasite-induced knobs on the infected erythrocyte surface, which are electron-dense features comprising many parasite proteins anchored to the erythrocyte cytoskeleton. Failure to present PfEMP1 in such knob structures greatly reduces the ability of the infected erythrocyte to bind to its specific host receptor [79].

Pf *var* gene regulation is complex and includes mechanisms of gene regulation such as chromosomal organization and subnuclear compartmentalization [80, 81], endogenous *var* gene clustering and *var* promoter–intron pairing [82, 83], transcriptional gene silencing via exoribonuclease-mediated RNA degradation [84], histone variant exchange at *var* promoters [85, 86], the effect of *trans* antisense long non-coding RNAs (lncRNAs) [87], and the presence or absence of histone modifications and their associated histone-modifying enzymes [27, 29, 40, 87–92] (Fig. 3c). Interest in delineating these mechanisms has continued, and even grown, as more research in the post-genomic area has highlighted the important differential role of the 5′ upstream promoter families into which the *var* genes can be subdivided into five classes (*upsA* to *upsE*), which correlate closely with the severity of malaria infection in the human host [93–98]. Pf *var* gene promoters are also essential components of the gene silencing mechanism and monoallelic expression. The *upsC var* promoter in particular is necessary to maintain chromosome-internal *var* genes in their silenced state and recently has been proposed to do so through the interaction of *cis*-acting MEE2-like sequence motifs and MEE2-interacting factors to reinforce *var* gene transcriptional repression [75, 83].

Monoallelic *var* gene transcription is also associated with the presence of H3K9me3 repressive marks at silent *var* gene loci (Fig. 3c). This histone modification is predicted but not proven to be imposed by the HKMT PfSET3 and is associated with perinuclear repressive centers and the binding of PfHP1, stimulating heterochromatin formation [40, 89, 90, 92]. Conditional disruption of one of these essential proteins, HP1, disrupts singular *var* gene expression and dysregulates antigenic variation [29]. In addition, conditional knockdown of PfHDA2 has been shown to result in a dramatic loss of monoallelic *var* gene expression [27]. This implicated PfHDA2 as an upstream regulator of HP1 binding as it facilitates the establishment of the H3K9me3 mark. The indispensable role of the dynamic histone lysine methylation of *Plasmodium* chromatin by histone lysine demethylases (HKDMs) and HKMTs in controlling the transcription of nearly all *var* genes has also been demonstrated. Knockout of the Pf *hkmt* gene encoding SET2/SETvs (vs, variant-silencing) resulted in reduced presence of the repressive H3K36me3 mark at the TSSs and intronic promoters of all *var* gene subtypes (Fig. 3c). Loss of this SETvs-dependent histone modification resulted in the loss of monoallelic *var* gene expression and expression of the entire *var* repertoire [98]. Furthermore, SETvs can directly interact with the C-terminal domain of RNA polymerase II, with SETvs disruption resulting in a loss of binding to RNA polymerase II and *var* gene switching [99].

Pf *upsA*-type *var* gene expression is also regulated by PfRNaseII, a chromatin-associated exoribonuclease. An inverse relationship exists between transcript levels of PfRNaseII and *upsA*-type *var* genes, with an increase in the latter corresponding to incidences of severe malaria in infected patients [84]. PfRNaseII is proposed to control *upsA*-type *var* gene transcription by marking TSSs and intronic promoter regions, degrading potential full-length transcripts to produce short-lived cryptic RNA molecules that are then further degraded by the exosome immediately upon expression (Fig. 3c). Disruption of the *pfrnaseII* gene resulted in loss of this degradation and the generation of full-length *upsA var* gene transcripts and intron-derived antisense lncRNA. These data illustrate the relationship between PfRNaseII and the control of monoallelic *var* gene transcription and suggest a correlation between lncRNA and *var* gene activation in Pf [84]. The role of lncRNAs in Pf *var* gene activation was again investigated in a study by Amit-Avraham and colleagues [87], which demonstrated dose-dependent transcriptional activation of *var* genes by overexpression of their individual antisense lncRNA transcripts. Disruption of antisense lncRNA expression by peptic nucleic acids resulted in downregulation of active *var* gene transcripts and induced *var* gene switching. The exact mechanism by which antisense lncRNAs act to promote the active transcription of a *var* gene is unknown. It has been postulated that antisense *var* transcripts may recruit chromatin-modifying enzymes that in turn would affect gene accessibility for the Pf transcriptional machinery. Antisense *var* gene lncRNAs would also contain a complementary sequence to *var* gene intronic insulator-like pairing elements that bind specific nuclear-binding proteins, therefore blocking the silencing activity of pairing elements by hybridization [87, 100].

The *Plasmodium* helical interspersed subtelomeric protein (PHIST) family of genes, which is unique to Pf, has also been implicated in the regulation of immune evasion as a result of its ability to bind to the intracellular acidic terminal segment of PfEMP1. Conditional knockdown of the essential PHIST protein PfE1605w reduced the capability of the infected host erythrocyte to adhere to the CD36 endothelial receptor, an important virulence feature of Pf. This study highlighted the importance not only of *var* genes and their controlled expression but also of other genes that are associated with anchoring PfEMP1 to the erythrocyte surface and creating the *Plasmodium* cytoadherence complex [101].

The list of regulatory mechanisms underlying *var* gene monoallelic expression is vast and much more may still be discovered in this area. However, immune evasion in the *Plasmodium* genus is not confined to Pf or *var* gene regulation. Indeed, *var* gene expression is exclusive to Pf, with much still to be gleaned in the areas of immune evasion in human malaria parasites such as *P. vivax*, *P. knowlesi*, *Plasmodium ovale* and *Plasmodium malariae* [13, 102–105]. In addition, PfEMP1 is just one of a number of variant surface antigens (VSAs) known to be expressed at the host erythrocyte surface upon infection with Pf, although it is the best characterized. Pf-infected erythrocytes also express VSAs of the multi-copy gene families of the proteins repetitive interspersed family (RIFIN), subtelomeric variable open reading frame (STEVOR), and Pf Maurer's cleft 2 transmembrane (PfMC-2TM) [106]. The roles of these protein families in antigenic variation and pathology are generally poorly defined but are being elucidated; for example, RIFINs are implicated in the severity of Pf malaria in African children with blood group A. This tendency toward increased malaria pathogenicity is a result of their expression at the surface of infected host erythrocytes, from which they bind uninfected erythrocytes (preferentially, blood group A) to form rosette structures and mediate binding to the host microvasculature [107]. Thus, the combined roles of HP1 and HDA2 in regulating single *var* gene expression and the transcriptional regulator ApiAP2-G suggests that both processes share epigenetic regulatory mechanisms and that *Plasmodium* immune evasion and transmission to new hosts are inextricably linked [27, 29].

Immune evasion is not restricted to blood-stage *Plasmodium*; when the parasite passes through the mosquito it must also combat a sophisticated innate immune system that is very effective in reducing the parasite burden experienced by the vector. A forward genetic screen and WGS was used to identify the key parasite factor, the surface protein PfS47 (found on the surface of the ookinete as it penetrates the mosquito midgut), that appears to interact with and suppress the vector innate immune system [108]. PfS47 is thought to suppress signaling through the c-Jun N-terminal kinase (JNK) pathway that is critical to an effective immune response [109]. WGS demonstrated that PfS47 has a distinct population structure linked to global distribution. PfS47 is rapidly evolving and selected to achieve JNK suppression in diverse mosquito species, which becomes a key step in the adaptation of Pf to transmission in different vectors, thereby contributing to its broad global distribution [110].

Artemisinin resistance

The goals of MalariaGEN characterize a new approach to understanding parasite population biology. Through the generation and, these days more critically, the management and analysis of the colossal datasets that result from WGS of large numbers of samples, a well-organized study can draw meaningful conclusions. This was applied to perhaps the most serious threat to malaria control that has emerged in recent years—resistance to ART. Using these datasets in meta-analyses with clinical data describing the individual WGS-sequenced samples and outcomes of ART treatment allowed a path to be charted that associated SNPs with treatment features (such as delayed clearance) [111] and identified candidate genes [112]: in both studies a region of chromosome 13 was implicated (Fig. 3d). The precise gene encoding the protein KELCH13 was identified by a combination of "old-fashioned" selection of drug-resistant parasites in the laboratory followed by WGS and comparative genomics of the sensitive parental parasites and the progeny, as well as WGS of ART-resistant field isolates [113, 114]. The role of the *kelch13* mutations in ART resistance was proven by direct genome engineering of *kelch13* to generate resistant parasites [115, 116]. *kelch13* SNPs have been used to map the alarmingly rapid spread of resistance throughout Southeast Asia [116] and it is clear that there is already significant but distinct *kelch13* heterogeneity in African Pf strains, although there is no evidence of ART resistance [117–121]. However, in-depth analysis of Southeast Asian ART-resistant parasite genomes [122] revealed that a complex array of background mutations (Fig. 3d) in a variety of genes (encoding ferredoxin (FD), apicoplast ribosomal protein S10 (ARPS10), multidrug resistance protein 2 (MDR2), and Pf chloroquine resistance transporter (CRT)) that are not yet described in African parasites would explain why ART resistance is not (yet) a threat to the use of ART in that continent [121].

A further puzzle was the large number of independent SNPs that seemed capable of mediating ART resistance—typically drug resistance is generated by one or a small number of SNPs focused on either altering the target binding site for the drug or preventing drug access to a binding site buried in the target structure. KELCH proteins are propeller proteins with an iterated structural motif that serves as a platform for the assembly of multiprotein complexes. In addition, KELCH13 has a BTB/POZ domain that might be involved in homodimerization, E3 ubiquitin ligase binding, and transcriptional repression (reviewed in [123]). It has been suggested that ART-resistance-associated *kelch13* SNPs might cause a degree of reduced binding of Pf phosphatidylinositol-3-kinase (PI3K), which in turn results in its reduced ubiquitination and consequent degradation of PI3K (Fig. 3d). Elevated levels of PI3K generate increased amounts of its lipid product phosphatidylinositol-3-phosphate (PI3P), which then changes the physiological state of the parasite cell through signaling in as yet unknown pathways [124] but through a mechanism predicated on the proposed abundance of PI3P in the lumen of the endoplasmic reticulum and its proposed role in protein export beyond the parasite vacuole within the host cell [125]. However, aspects of this view have been challenged [126] and further studies are clearly required to resolve the possible role of PI3K signaling in ART resistance. It will be of interest to see if PI3K signaling impacts upon the unfolded protein response implicated in ART resistance using population transcriptomics [127]. The WGS data and two proteomics studies [128, 129] that demonstrate the wide variety of proteins from different cellular compartments of the target parasite that interact with activated ART together suggest that ART resistance is a pleiotropic phenomenon [123]. Therefore, other interrogations, such as metabolomics (see next section), might also be needed to gain functional insights into ART's mode of action.

Translational implications for malaria control

Antimalarials

WGS has been instrumental in identifying the cellular target of novel Pf antimalarials as part of the drug discovery pipeline and in following the in vitro selection of resistant parasite lines and validation of observed genomic changes by reverse genetics as described for ART above. This approach has proved highly successful for spiralindolines [130], resulting in the identification of the target of NITD609 (also known as KAE609 or cipargamin) as the P-type ATPase PfATPase4. Furthermore, the translation elongation factor eEF2 has been identified as the target of the 2,6-disubstituted quinoline-4-carboxamide scaffold derivative DDD107498 [131]. WGS is not the only post-genome approach that is useful for determining modes of drug action; metabolomics has a similar potential for analyzing the metabolic changes produced in response to drug exposure and has been utilized in antibiotic [132] and anti-protozoan drug [133] investigations. A metabolomics-based approach also has the advantages that parasite lines resistant to the drug need not be generated and that the activity of pleiotropically acting drugs (such as ART) are directly observed rather than imputed from genomes of resistant parasites.

Vaccines

Post-genomic approaches have also identified promising new Pf vaccine candidates. For example, Pf reticulocyte-binding protein homologue 5 (RH5) binds to the human red cell surface receptor protein basigin, an interaction that is essential for erythrocyte invasion by Pf [134]. Recent WGS studies have shown that both the host and parasite proteins are highly conserved, that antibodies to RH5 block merozoite invasion of erythrocytes [135, 136], and that basigin itself is druggable by recombinant antibodies [137]. Although RH5–basigin interaction offers great promise, the challenges for vaccine development remain considerable and many promising candidates have fallen or will fall by the wayside due to an inability to formulate them to deliver effective vaccination, massive candidate gene sequence variability, and functional non-essentiality of the candidate. WGS will help identify non- or minimally variant candidates and should prove useful in monitoring the effect of vaccination and the analysis of "breakthrough" parasites (those developing in vaccinated individuals), as described in the next section. Effective subunit vaccines will be an invaluable additional approach to vaccination, supplementing other approaches such as the use of the promising but technologically challenging attenuated whole parasite, for example, sporozoite vaccine [138].

Surveillance

The identification of genome signatures of resistance through WGS in the laboratory and increasingly through large-scale genomic epidemiology provides a powerful tool to monitor the emergence of resistance in *Plasmodium* populations under selective pressure due to administration of both drugs and vaccines. In the case of drugs whose targets have been identified in the laboratory, specific, simple PCR-based assays can be devised. WGS of field parasites under drug pressure is still desirable, however, as alternative resistance mechanisms might emerge that would be missed by targeted assays and, with sufficient depth of sampling, new signatures of resistance would be identifiable from the sequence data. Similar surveillance of parasites that emerge post-vaccination might also be informative. An important analysis of the clinical trial of the RTS,S/AS01 malaria vaccine compared the strain-specific sequence of the gene encoding the circumsporozoite (CS) protein that comprised the vaccine with the CS gene sequences of strains in the infections actually encountered by immunized individuals (between 5 and 17 months of age) [139]. This study demonstrated that homologous protection was greater than protection against heterologous strains and that a cause of failure to protect was simply that the CS protein carried by the infecting parasites did not match that of the vaccine and so perhaps a protective effect was less likely [139]. Therefore, WGS has the power to guide vaccine design based upon the outcomes of trials.

Gene editing

A new era of genetic engineering has dawned with the discovery and development of the bacterial guide RNA template-targeted clustered regularly interspaced short palindromic repeats (CRISPR)-Cas9 recombinases as tools for the accurate editing of genomes. The technology has been successfully adapted to many species, including *Plasmodium* [140], *Anopheles* [141, 142], and humans (discussed in [143]). Currently, applications of CRISPR-Cas9 to *Plasmodium* manipulation are restricted to reverse genetic investigations of gene function. However, with the concepts of whole (pre-erythrocytic) parasite vaccines [144, 145], CRISPR-Cas9 offers an obvious route towards the generation of an immunogenic, non-pathogenic parasite that might be suitably safe to administer to humans as a vaccination strategy. Clearly, the engineering of human genomes at any stage of gestation is fraught with ethical considerations [146] and it is inconceivable that this will be applied to the improvement of human resistance to malaria in the foreseeable future. Conversely, although subject to similar ethical and ecological debate, significant conceptual advances towards the generation of CRISPR-Cas9-engineered *Anopheles* mosquitoes have been quickly achieved. Through harnessing the concept of gene drive, two independent teams have reported the generation of either engineered *Anopheles stephensi* (a major Indian vector of malaria) that is resistant to malaria [141] or sterile female Ag [142]. Again, owing to ecological considerations, it is unlikely that such engineered mosquitoes, although clearly feasible, will be released into the wild any time soon [147].

Conclusions and future directions

Despite the progress summarized here, the fundamental requirements of malaria research in any era remain the same; namely, new drugs to replace those that become ineffective, vaccines that work, and the means to administer them effectively. Genomics, post-genomic technologies, and associated computing developments have revolutionized investigations into the biology of the malaria parasite and the search for therapeutics or intervention measures. Significant progress has been made on many fronts, including candidate drug and vaccine discovery, parasite drug resistance mechanisms, host–parasite–vector interactions and parasite biology, and mechanisms of human resistance to malaria. Also, new concepts of combatting malaria via engineered mosquito populations have been introduced through the advent of novel genome-editing approaches such as CRISPR-Cas9.

We can anticipate that WGS will continue to improve in terms of both cost and quality, making sequencing of every desirable Pf isolate feasible. This would enable more detailed studies of population structure and dynamics, allowing the tracking of gene flow and genotype success that might even resolve at the village level and, further, potentially almost in real time. However, this will only happen if data storage, access, and computing technologies keep pace. Where Pf WGS studies have gone *P. vivax* research will follow and recent studies have revealed signatures of drug selection superimposed upon a far more complex (global, regional, and even within a single infection) population structure than Pf [148, 149]. Single-cell RNA sequencing will significantly improve our understanding of antigenic variation and variant and sex-specific gene expression.

More immediately, an important need is for surveillance, particularly in Africa, to look for *kelch13* mutations and genotypes associated with ART resistance and a pan-African network is in place to monitor for this and collect samples [150]. Genomics will continue to be used in novel ways as well, for example, in studies of the outcomes of human interventions such as drug treatment and vaccination.

New fields of endeavor are also emerging that will certainly prove fruitful in the years to come. Lipidomics is a nascent discipline that will no doubt reveal insights into membrane composition and organization [151] and might also open avenues to therapy. PTMs such as palmitoylation give proteins the means to conditionally interact with membranes and *Plasmodium* makes extensive use of protein palmitoylation that should influence a range of important parasite biological activities, such as cytoadherence and drug resistance [152].

Although the power of genomics approaches is quite clear, direct biological investigations are frequently required to confirm or refute the findings that genomics might imply. The numerous examples given here indicate that although genomic analyses often generate associations and degrees of confidence about their conclusions, unequivocal confirmation is provided by genetic engineering (of parasites and their vectors at least). Genetic screens are powerful, often unbiased approaches to discover gene function. The recent development of the PlasmoGEM resource coupled with high-efficiency transfection and barcoded vectors permits genome-scale reverse genetics screens to be deployed that will undoubtedly reveal information about parasite-specific genes and *Plasmodium* biology [153]. Finally, many of the genes encoded by parasite, host, and vector genomes have unknown functions, the details of which are slowly emerging as technologies and assays improve. The staggering complexity of organismal biology and the interactions between parasite, host, and vector will continue to amaze but equally will offer hope for new and improved therapies.

Abbreviations

(l)ncRNA, (Long) Non-coding RNA; Ag, *Anopheles Gambiae*; AID, Auxin-inducible Degron; ART, Artemisinin; circRNA, Circular RNAs; cKD, Conditional Knockdown; CRISPR, Clustered Regularly Interspaced Short Palindromic Repeats; DD, Destabilization Domain; G6PD, Glucose-6-phosphate Dehydrogenase; GWAS, Genome-wide Association Study; IDC, Intra-erythrocytic Development Cycle; K13, Kelch13; MalariaGEN, Malaria Genomic Epidemiology Network; miRNA, MicroRNA; NGS, Next-generation Sequencing; ORF, Open Reading Frame; Pb, Plasmodium Berghei; Pf, Plasmodium Falciparum; PHIST, Plasmodium Helical Interspersed Subtelomeric Protein Family; PI3K, Phosphatidylinositol-3-kinase; PI3P, Phosphatidylinositol-3-phosphate; PTM, Post-translational Modification; RBC, Red Blood Cell; rDNA, Ribosomal DNA; RIFIN, Repetitive Interspersed Family; SNP, Single-nucleotide Polymorphism; TF, Transcription Factor; TSS, Transcription Start Site; UPR, Unfolded Protein Response; VSAs, Variant Surface Antigens; WGS, Whole-genome Sequencing; ZFN, Zinc Finger Nuclease

Acknowledgements

SK is a Marie Skłodowska-Curie Fellow (reference 657592). BJP is funded by the Wellcome Trust (reference 102463/Z/13/Z). APW is funded by the Wellcome Trust (reference 083811/Z/07/Z).

Authors' contributions

All authors wrote and proofed the entire manuscript and then read and approved the final manuscript. APW managed the writing and organized the article. SK produced all figures and tables after mutual discussion and agreement about the format and contents.

Competing interests

The authors declare that they have no competing interests.

Published online: 07 September 2016

References

1. World Health Organization. World Malaria Report 2015. 2015. http://apps.who.int/iris/bitstream/10665/200018/1/9789241565158_eng.pdf?ua=1. Accessed 24 June 2016.
2. Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, et al. Genome sequence of the human malaria parasite *Plasmodium falciparum*. Nature. 2002;419:498–511.
3. Larremore DB, Sundararaman SA, Liu W, Proto WR, Clauset A, Loy DE, et al. Ape parasite origins of human malaria virulence genes. Nat Commun. 2015;6:8368.
4. Kwiatkowski D. Malaria genomics: tracking a diverse and evolving parasite population. Int Health. 2015;7:82–4.
5. Malaria Genomic Epidemiology Network. A novel locus of resistance to severe malaria in a region of ancient balancing selection. Nature. 2015;526:253–7.
6. Redmond SN, Eiglmeier K, Mitri C, Markianos K, Guelbeogo WM, Gneme A, et al. Association mapping by pooled sequencing identifies TOLL 11 as a protective factor against *Plasmodium falciparum* in *Anopheles gambiae*. BMC Genomics. 2015;16:779.
7. Gurdasani D, Carstensen T, Tekola-Ayele F, Pagani L, Tachmazidou I, Hatzikotoulas K, et al. The African Genome Variation Project shapes medical genetics in Africa. Nature. 2015;517:327–32.
8. Malaria Genomic Epidemiology Network. Reappraisal of known malaria resistance loci in a large multicenter study. Nat Genet. 2014;46:1197–204.
9. Maier AG, Duraisingh MT, Reeder JC, Patel SS, Kazura JW, Zimmerman PA, et al. *Plasmodium falciparum* erythrocyte invasion through glycophorin C and selection for Gerbich negativity in human populations. Nat Med. 2003;9:87–92.
10. Fontaine MC, Pease JB, Steele A, Waterhouse RM, Neafsey DE, Sharakhov IV, et al. Extensive introgression in a malaria vector species complex revealed by phylogenomics. Science. 2015;347:1258524.
11. Neafsey DE, Waterhouse RM, Abai MR, Aganezov SS, Alekseyev MA, Allen JE, et al. Highly evolvable malaria vectors: the genomes of 16 *Anopheles* mosquitoes. Science. 2015;347:1258522.
12. Carlton JM, Adams JH, Silva JC, Bidwell SL, Lorenzi H, Caler E, et al. Comparative genomics of the neglected human malaria parasite *Plasmodium vivax*. Nature. 2008;455:757–63.
13. Pain A, Böhme U, Berry AE, Mungall K, Finn RD, Jackson AP, et al. The genome of the simian and human malaria parasite *Plasmodium knowlesi*. Nature. 2008;455:799–803.
14. Keeling PJ, Rayner JC. The origins of malaria: there are more things in heaven and earth. Parasitology. 2015;142 Suppl 1:S16–25.
15. Liu W, Sundararaman SA, Loy DE, Learn GH, Li Y, Plenderleith LJ, et al. Multigenomic delineation of *Plasmodium* species of the *Laverania* subgenus infecting wild-living chimpanzees and gorillas. Genome Biol Evol. 2016;8:1929–39.
16. Mourier T, Carret C, Kyes S, Christodoulou Z, Gardner PP, Jeffares DC, et al. Genome-wide discovery and verification of novel structured RNAs in *Plasmodium falciparum*. Genome Res. 2008;18:281–92.
17. Broadbent KM, Broadbent JC, Ribacke U, Wirth D, Rinn JL, Sabeti PC. Strand-specific RNA sequencing in *Plasmodium falciparum* malaria identifies developmentally regulated long non-coding RNA and circular RNA. BMC Genomics. 2015;16:454.
18. Siegel TN, Hon CC, Zhang Q, Lopez-Rubio JJ, Scheidig-Benatar C, Martins RM, et al. Strand-specific RNA-Seq reveals widespread and developmentally regulated transcription of natural antisense transcripts in *Plasmodium falciparum*. BMC Genomics. 2014;15:150.
19. Manske M, Miotto O, Campino S, Auburn S, Almagro-Garcia J, Maslen G, et al. Analysis of *Plasmodium falciparum* diversity in natural infections by deep sequencing. Nature. 2012;487:375–9.
20. Bonasio R, Tu S, Reinberg D. Molecular signals of epigenetic states. Science. 2010;330:612–6.
21. Cui L, Miao J. Chromatin-mediated epigenetic regulation in the malaria parasite *Plasmodium falciparum*. Eukaryot Cell. 2010;9:1138–49.
22. Saraf A, Cervantes S, Bunnik EM, Ponts N, Sardiu ME, Chung DD, et al. Dynamic and combinatorial landscape of histone modifications during the intraerythrocytic developmental cycle of the malaria parasite. J Proteome Res. 2016; doi:10.1021/acs.jproteome.6b00366.
23. Volz J, Carvalho TG, Ralph SA, Gilson P, Thompson J, Tonkin CJ, et al. Potential epigenetic regulatory proteins localise to distinct nuclear sub-compartments in *Plasmodium falciparum*. Int J Parasitol. 2010;40:109–21.
24. Doerig C, Rayner JC, Scherf A, Tobin AB. Post-translational protein modifications in malaria parasites. Nat Rev Microbiol. 2015;13:160–72.
25. Chen PB, Ding S, Zanghì G, Soulard V, Dimaggio PA, Fuchter MJ, et al. *Plasmodium falciparum Pf*SET7: enzymatic characterization and cellular localization of a novel protein methyltransferase in sporozoite, liver and erythrocytic stage parasites. Sci Rep. 2016;6:21802.
26. Biamonte MA, Wanner J, Le Roch KG. Recent advances in malaria drug discovery. Bioorg Med Chem Lett. 2013;23:2829–43.
27. Coleman BI, Skillman KM, Jiang RH, Childs LM, Altenhofen LM, Ganter M, et al. A *Plasmodium falciparum* histone deacetylase regulates antigenic variation and gametocyte conversion. Cell Host Microbe. 2014;16:177–86.
28. Bártfai R, Hoeijmakers WA, Salcedo-Amaya AM, Smits AH, Janssen-Megens E, Kaan A, et al. H2A.Z demarcates intergenic regions of the *plasmodium falciparum* epigenome that are dynamically marked by H3K9ac and H3K4me3. PLoS Pathog. 2010;6:e1001223.
29. Brancucci NM, Bertschi NL, Zhu L, Niederwieser I, Chin WH, Wampfler R, et al. Heterochromatin protein 1 secures survival and transmission of malaria parasites. Cell Host Microbe. 2014;16:165–76.
30. Lomberk G, Wallrath L, Urrutia R. The heterochromatin protein 1 family. Genome Biol. 2006;7:228.
31. Josling GA, Petter M, Oehring SC, Gupta AP, Dietz O, Wilson DW, et al. A *Plasmodium falciparum* bromodomain protein regulates invasion gene expression. Cell Host Microbe. 2015;17:741–51.
32. Bunnik EM, Polishko A, Prudhomme J, Ponts N, Gill SS, Lonardi S, et al. DNA-encoded nucleosome occupancy is associated with transcription levels in the human malaria parasite *Plasmodium falciparum*. BMC Genomics. 2014;15:347.
33. Ponts N, Harris EY, Prudhomme J, Wick I, Eckhardt-Ludka C, Hicks GR, et al. Nucleosome landscape and control of transcription in the human malaria parasite. Genome Res. 2010;20:228–38.

34. Kensche PR, Hoeijmakers WAM, Toenhake CG, Bras M, Chappell L, Berriman M, et al. The nucleosome landscape of *Plasmodium falciparum* reveals chromatin architecture and dynamics of regulatory sequences. Nucleic Acids Res. 2015. doi:10.1093/nar/gkv1214.
35. Hoeijmakers WA, Salcedo-Amaya AM, Smits AH, Françoijs KJ, Treeck M, Gilberger TW, et al. H2A.Z/H2B.Z double-variant nucleosomes inhabit the AT-rich promoter regions of the *Plasmodium falciparum* genome. Mol Microbiol. 2013;87:1061–73.
36. Pombo A, Dillon N. Three-dimensional genome architecture: players and mechanisms. Nat Rev Mol Cell Biol. 2015;16:245–57.
37. Ay F, Bunnik EM, Varoquaux N, Bol SM, Prudhomme J, Vert JP, et al. Three-dimensional modeling of the *P. falciparum* genome during the erythrocytic cycle reveals a strong connection between genome architecture and gene expression. Genome Res. 2014;24:974–88.
38. de Wit E, de Laat W. A decade of 3C technologies: insight into nuclear organization. Genes Dev. 2012;26:11–24.
39. Lemieux JE, Kyes SA, Otto TD, Feller AI, Eastman RT, Pinches RA, et al. Genome-wide profiling of chromosome interactions in *Plasmodium falciparum* characterizes nuclear architecture and reconfigurations associated with antigenic variation. Mol Microbiol. 2013;90:519–37.
40. Lopez-Rubio JJ, Mancio-Silva L, Scherf A. Genome-wide analysis of heterochromatin associates clonally variant gene regulation with perinuclear repressive centers in malaria parasites. Cell Host Microbe. 2009;5:179–90.
41. Vembar SS, Droll D, Scherf A. Translational regulation in blood stages of the malaria parasite *Plasmodium* spp.: systems-wide studies pave the way. Wiley Interdiscip Rev RNA. 2016. doi:10.1002/wrna.1365.
42. Adjalley SH, Chabbert CD, Klaus B, Pelechano V, Steinmetz LM. Landscape and dynamics of transcription initiation in the malaria parasite *Plasmodium falciparum*. Cell Rep. 2016;14:2463–75.
43. Balaji S, Babu MM, Iyer LM, Aravind L. Discovery of the principal specific transcription factors of Apicomplexa and their implication for the evolution of the AP2-integrase DNA binding domains. Nucleic Acids Res. 2005;33:3994–4006.
44. Painter HJ, Campbell TL, Llinás M. The Apicomplexan AP2 family: integral factors regulating *Plasmodium* development. Mol Biochem Parasitol. 2011;176:1–7.
45. Iwanaga S, Kaneko I, Kato T, Yuda M. Identification of an AP2-family protein that is critical for malaria liver stage development. PLoS One. 2012;7, e47557.
46. Yuda M, Iwanaga S, Shigenobu S, Mair GR, Janse CJ, Waters AP, et al. Identification of a transcription factor in the mosquito-invasive stage of malaria parasites. Mol Microbiol. 2009;71:1402–14.
47. Yuda M, Iwanaga S, Shigenobu S, Kato T, Kaneko I. Transcription factor AP2-Sp and its target genes in malarial sporozoites. Mol Microbiol. 2010;75:854–63.
48. Kaneko I, Iwanaga S, Kato T, Kobayashi I, Yuda M. Genome-wide identification of the target genes of AP2-O, a *Plasmodium* AP2-family transcription factor. PLoS Pathog. 2015;11, e1004905.
49. Kafsack BF, Rovira-Graells N, Clark TG, Bancells C, Crowley VM, Campino SG, et al. A transcriptional switch underlies commitment to sexual development in malaria parasites. Nature. 2014;507:248–52.
50. Sinha A, Hughes KR, Modrzynska KK, Otto TD, Pfander C, Dickens NJ, et al. A cascade of DNA-binding proteins for sexual commitment and development in *Plasmodium*. Nature. 2014;507:253–7.
51. Yuda M, Iwanaga S, Kaneko I, Kato T. Global transcriptional repression: an initial and essential step for *Plasmodium sexual* development. Proc Natl Acad Sci U S A. 2015;112:12824–9.
52. Waters AP. Epigenetic roulette in blood stream *Plasmodium*: gambling on sex. PLoS Path. 2016;12, e1005353.
53. Rovira-Graells N, Gupta AP, Planet E, Crowley VM, Mok S, Ribas de Pouplana L, et al. Transcriptional variation in the malaria parasite *Plasmodium falciparum*. Genome Res. 2012;22:925–38.
54. Miao J, Lawrence M, Jeffers V, Zhao F, Parker D, Ge Y, et al. Extensive lysine acetylation occurs in evolutionarily conserved metabolic pathways and parasite-specific functions during *Plasmodium falciparum* intraerythrocytic development. Mol Microbiol. 2013;89:660–75.
55. Cobbold SA, Santos JM, Ochoa A, Perlman DH, Llinás M. Proteome-wide analysis reveals widespread lysine acetylation of major protein complexes in the malaria parasite. Sci Rep. 2016;6:19722.
56. Jackson KE, Habib S, Frugier M, Hoen R, Khan S, Pham JS, et al. Protein translation in *Plasmodium* parasites. Trends Parasitol. 2011;27:467–76.
57. Le Roch KG, Zhou Y, Blair PL, Grainger M, Moch JK, Haynes JD, et al. Discovery of gene function by expression profiling of the malaria parasite life cycle. Science. 2003;301:1503–08.
58. Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science. 2009;324:218–23.
59. Caro F, Ahyong V, Betegon M, DeRisi JL. Genome-wide regulatory dynamics of translation in the *Plasmodium falciparum* asexual blood stages. Elife. 2014;10:3.
60. Mair GR, Braks JA, Garver LS, Wiegant JC, Hall N, Dirks RW, et al. Regulation of sexual development of *Plasmodium* by translational repression. Science. 2006;313:667–9.
61. Silva PA, Guerreiro A, Santos JM, Braks JA, Janse CJ, Mair GR. Translational control of UIS4 protein of the host-parasite interface is mediated by the RNA binding protein Puf2 in *Plasmodium berghei* sporozoites. PLoS One. 2016;11, e0147940.
62. Gomes-Santos CSS, Brak JAM, Prudêncio M, Carret CK, Gomes ARB, Pain A, et al. *Plasmodium* sporozoites in the salivary gland are latent liver stage forms regulated by the RNA binding protein pumilio. PloS Path. 2011;7, e1002046.
63. Lasonder E, Rijpma SR, Van Schaijk BC, Hoeijmakers WA, Kensche PR, Gresnigt MS, et al. Integrated transcriptomic and proteomic analyses of *P. falciparum* gametocytes: molecular insight into sex-specific processes and translational repression. Nucleic Acids Res. 2016. doi:10.1093/nar/gkw536.
64. Guerreiro A, Deligianni E, Santos JM, Silva PA, Louis C, Pain A, et al. Genome-wide RIP-Chip analysis of translational repressor-bound mRNAs in the *Plasmodium* gametocyte. Genome Biol. 2014;15:493.
65. Vembar SS, Macpherson CR, Sismeiro O, Coppée JY, Scherf A. The *Pf*Alba1 RNA-binding protein is an important regulator of translational timing in *Plasmodium falciparum* blood stages. Genome Biol. 2015;16:212.
66. Bunnik EM, Chung DW, Hamilton M, Ponts N, Saraf A, Prudhomme J, et al. Polysome profiling reveals translational control of gene expression in the human malaria parasite *Plasmodium falciparum*. Genome Biol. 2013;14:R128.
67. Brancucci NM, Witmer K, Schmid C, Voss TS. A *var* gene upstream element controls protein synthesis at the level of translation initiation in *Plasmodium falciparum*. PLoS One. 2014;9, e100183.
68. Kumar M, Srinivas V, Patankar S. Upstream AUGs and upstream ORFs can regulate the downstream ORF in *Plasmodium falciparum*. Malar J. 2015;14:512.
69. Chakrabarti K, Pearson M, Grate L, Sterne-Weiler T, Deans J, Donohue JP, et al. Structural RNAs of known and unknown function identified in malaria parasites by comparative genomics and RNA analysis. RNA. 2007;13:1923–39.
70. Vembar SS, Scherf A, Siegel TN. Noncoding RNAs as emerging regulators of *Plasmodium falciparum* virulence gene expression. Curr Opin Microbiol. 2014;20:153–61.
71. LaMonte G, Philip N, Reardon J, Lacsina JR, Majoros W, Chapman L, et al. Translocation of sickle cell erythrocyte microRNAs into *Plasmodium falciparum* inhibits parasite translation and contributes to malaria resistance. Cell Host Microbe. 2012;12:187–99.
72. Chen LL. The biogenesis and emerging roles of circular RNAs. Nat Rev Mol Cell Biol. 2016;17:205–11.
73. Kwiatkowski DP. How malaria has affected the human genome and what human genetics can teach us about malaria. Am J Hum Genet. 2005;77:171–92.
74. Ferreira MU, Da Silva NM, Wunderlich G. Antigenic diversity and immune evasion by malaria parasites. Clin Diagn Lab Immunol. 2004;11:987–95.
75. Brancucci NM, Witmer K, Schmid CD, Flueck C, Voss TS. Identification of a *cis*-acting DNA-protein interaction implicated in singular *var* gene choice in *Plasmodium falciparum*. Cell Microbiol. 2012;14:1836–48.
76. Smith JD, Chitnis CE, Craig AG, Roberts DJ, Hudson-Taylor DE, Peterson DS, et al. Switches in expression of *Plasmodium falciparum var* genes correlate with changes in antigenic and cytoadherent phenotypes of infected erythrocytes. Cell. 1995;82:101–10.
77. Recker M, Buckee CO, Serazin A, Kyes S, Pinches R, Christodoulou Z, et al. Antigenic variation in *Plasmodium falciparum* malaria involves a highly structured switching pattern. PLoS Pathog. 2011;7, e1001306.
78. Almelli T, Ndam NT, Ezimegnon S, Alao MJ, Ahouansou C, Sagbo G, et al. Cytoadherence phenotype of *Plasmodium falciparum*-infected erythrocytes is associated with specific pfemp-1 expression in parasites from children with cerebral malaria. Malar J. 2014;13:333.

79. Subramani R, Quadt K, Jeppesen AE, Hempel C, Petersen JE, Hassenkam T, et al. *Plasmodium falciparum*-infected erythrocyte knob density is linked to the *Pf*EMP1 variant expressed. MBio. 2015;6:e01456–15.
80. Thompson JK, Rubio JP, Caruana S, Brockman A, Wickham ME, Cowman AF. The chromosomal organization of the *Plasmodium falciparum var* gene family is conserved. Mol Biochem Parasitol. 1997;87:49–60.
81. Freitas-Junior LH, Bottius E, Pirrit LA, Deitsch KW, Scheidig C, Guinet F, et al. Frequent ectopic recombination of virulence factor genes in telomeric chromosome clusters of *P. falciparum*. Nature. 2000;407:1018–22.
82. Swamy L, Amulic B, Deitsch KW. *Plasmodium falciparum var* gene silencing is determined by *cis* DNA elements that form stable and heritable interactions. Eukaryot Cell. 2011;10:530–9.
83. Voss TS, Healer J, Marty AJ, Duffy MF, Thompson JK, Beeson JG, et al. A *var* gene promoter controls allelic exclusion of virulence genes in *Plasmodium falciparum* malaria. Nature. 2006;439:1004–8.
84. Zhang Q, Siegel TN, Martins RM, Wang F, Cao J, Gao Q, et al. Exonuclease-mediated degradation of nascent RNA silences genes linked to severe malaria. Nature. 2014;513:431–5.
85. Petter M, Lee CC, Byrne TJ, Boysen KE, Volz J, Ralph SA, et al. Expression of *P. falciparum var* genes involves exchange of the histone variant H2A.Z at the promoter. PLoS Pathog. 2011;7, e1001292.
86. Petter M, Selvarajah SA, Lee CC, Chin WH, Gupta AP, Bozdech Z, et al. H2A.Z and H2B.Z double-variant nucleosomes define intergenic regions and dynamically occupy *var* gene promoters in the malaria parasite *Plasmodium falciparum*. Mol Microbiol. 2013;87:1167–82.
87. Amit-Avraham I, Pozner G, Eshar S, Fastman Y, Kolevzon N, Yavin E, et al. Antisense long noncoding RNAs regulate *var* gene activation in the malaria parasite *Plasmodium falciparum*. Proc Natl Acad Sci U S A. 2015;112:E982–91.
88. Lopez-Rubio JJ, Gontijo AM, Nunes MC, Issar N, Hernandez Rivas R, Scherf A. 5′ flanking region of *var* genes nucleate histone modification patterns linked to phenotypic inheritance of virulence traits in malaria parasites. Mol Microbiol. 2007;66:1296–305.
89. Flueck C, Bartfai R, Volz J, Niederwieser I, Salcedo-Amaya AM, Alako BT, et al. *Plasmodium falciparum* heterochromatin protein 1 marks genomic loci linked to phenotypic variation of exported virulence factors. PLoS Pathog. 2009;5, e1000569.
90. Pérez-Toledo K, Rojas-Meza AP, Mancio-Silva L, Hernández-Cuevas NA, Delgadillo DM, Vargas M, et al. *Plasmodium falciparum* heterochromatin protein 1 binds to tri-methylated histone 3 lysine 9 and is linked to mutually exclusive expression of *var* genes. Nucleic Acids Res. 2009;37:2596–606.
91. Duraisingh MT, Voss TS, Marty AJ, Duffy MF, Good RT, Thompson JK, et al. Heterochromatin silencing and locus repositioning linked to regulation of virulence genes in *Plasmodium falciparum*. Cell. 2005;121:13–24.
92. Freitas-Junior LH, Hernandez-Rivas R, Ralph SA, Montiel-Condado D, Ruvalcaba-Salazar OK, Rojas-Meza AP. Telomeric heterochromatin propagation and histone acetylation control mutually exclusive expression of antigenic variation genes in malaria parasites. Cell. 2005;121:25–36.
93. Su XZ, Heatwole VM, Wertheimer SP, Guinet F, Herrfeldt JA, Peterson DS, et al. The large diverse gene family *var* encodes proteins involved in cytoadherence and antigenic variation of *Plasmodium falciparum*-infected erythrocytes. Cell. 1995;82:89–100.
94. Lavstsen T, Salanti A, Jensen AT, Arnot DE, Theander TG. Sub-grouping of *Plasmodium falciparum* 3D7 *var* genes based on sequence analysis of coding and non-coding regions. Malar J. 2003;2:27.
95. Kyriacou HM, Stone GN, Challis RJ, Raza A, Lyke KE, Thera MA, et al. Differential *var* gene transcription in *Plasmodium falciparum* isolates from patients with cerebral malaria compared to hyperparasitaemia. Mol Biochem Parasitol. 2006;150:211–8.
96. Kyes SA, Kraemer SM, Smith JD. Antigenic variation in *Plasmodium falciparum*: gene organization and regulation of the *var* multigene family. Eukaryot Cell. 2007;6:1511–20.
97. Trimnell AR, Kraemer SM, Mukherjee S, Phippard DJ, Janes JH, Flamoe E, et al. Global genetic diversity and evolution of *var* genes associated with placental and severe childhood malaria. Mol Biochem Parasitol. 2006;148:169–80.
98. Jiang L, Mu J, Zhang Q, Ni T, Srinivasan P, Rayavara K, et al. PfSETvs methylation of histone H3K36 represses virulence genes in *Plasmodium falciparum*. Nature. 2013;499:223–7.
99. Ukaegbu UE, Kishore SP, Kwiatkowski DL, Pandarinath C, Dahan-Pasternak N, Dzikowski R, et al. Recruitment of *Pf*SET2 by RNA polymerase II to variant antigen encoding loci contributes to antigenic variation in *P. falciparum*. Plos Pathog. 2014;10, e1003854.
100. Avraham I, Schreier J, Dzikowski R. Insulator-like pairing elements regulate silencing and mutually exclusive expression in the malaria parasite *Plasmodium falciparum*. Proc Natl Acad Sci U S A. 2012;109:E3678–86.
101. Oberli A, Zurbrügg L, Rusch S, Brand F, Butler ME, Day JL, et al. *Plasmodium falciparum* PHIST proteins contribute to cytoadherence and anchor *Pf*EMP1 to the host cell cytoskeleton. Cell Microbiol. 2016. doi:10.1111/cmi.12583.
102. Flick K, Chen Q. *var* genes, *Pf*EMP1 and the human host. Mol Biochem Parasitol. 2004;134:3–9.
103. Lapp SA, Korir-Morrison C, Jiang J, Bai Y, Corredor V, Galinski MR. Spleen-dependent regulation of antigenic variation in malaria parasites: *Plasmodium knowlesi* SICAvar expression profiles in splenic and asplenic hosts. PLoS One. 2013;8, e78014.
104. Neafsey DE, Galinsky K, Jiang RH, Young L, Sykes SM, Saif S, et al. The malaria parasite *Plasmodium vivax* exhibits greater genetic diversity than *Plasmodium falciparum*. Nat Genet. 2012;44:1046–50.
105. Fernandez-Becerra C, Yamamoto MM, Vêncio RZ, Lacerda M, Rosanas-Urgell A, Del Portillo HA. *Plasmodium vivax* and the importance of the subtelomeric multigene *vir* superfamily. Trends Parasitol. 2009;25:44–51.
106. Frech C, Chen N. Variant surface antigens of malaria parasites: functional and evolutionary insights from comparative gene family classification and analysis. BMC Genomics. 2013;14:427.
107. Bachmann A, Scholz JA, Janssen M, Klinkert MQ, Tannich E, Bruchhaus I, et al. A comparative study of the localization and membrane topology of members of the RIFIN, STEVOR and PfMC-2TM protein families in *Plasmodium falciparum*-infected erythrocytes. Malar J. 2015;14:274.
108. Molina-Cruz A, Garver LS, Alabaster A, Bangiolo L, Haile A, Winikor J, et al. The human malaria parasite Pfs47 gene mediates evasion of the mosquito immune system. Science. 2013;340:984–7.
109. Ramphul UN, Garver LS, Molina-Cruz A, Canepa GE, Barillas-Mury C. *Plasmodium falciparum* evades mosquito immunity by disrupting JNK-mediated apoptosis of invaded midgut cells. Proc Natl Acad Sci U S A. 2015;112:1273–80.
110. Molina-Cruz A, Canepa GE, Kamath N, Pavlovic NV, Mu J, Ramphul UN, et al. *Plasmodium* evasion of mosquito immunity and global malaria transmission: the lock-and-key theory. Proc Natl Acad Sci U S A. 2015;112:15178–83.
111. Takala-Harrison S, Clark TG, Jacob CG, Cummings MP, Miotto O, Dondorp AM, et al. Genetic loci associated with delayed clearance of *Plasmodium falciparum* following artemisinin treatment in Southeast Asia. Proc Natl Acad Sci U S A. 2013;110:240–5.
112. Cheeseman IH, Miller BA, Nair S, Nkhoma S, Tan A, Tan JC, et al. A major genome region underlying artemisinin resistance in malaria. Science. 2012;336:79–82.
113. Borrmann S, Straimer J, Mwai L, Abdi A, Rippert A, Okombo J, et al. Genome-wide screen identifies new candidate genes associated with artemisinin susceptibility in *Plasmodium falciparum* in Kenya. Sci Rep. 2013;3:3318.
114. Ariey F, Witkowski B, Amaratunga C, Beghain J, Langlois AC, Khim N, et al. A molecular marker of artemisinin-resistant *Plasmodium falciparum* malaria. Nature. 2014;505:50–5.
115. Ghorbal M, Gorman M, Macpherson CR, Martins RM, Scherf A, Lopez-Rubio JJ. Genome editing in the human malaria parasite *Plasmodium falciparum* using the CRISPR-Cas9 system. Nat Biotechnol. 2014;32:819–21.
116. Straimer J, Gnädig NF, Witkowski B, Amaratunga C, Duru V, Ramadani AP. K13-propeller mutations confer artemisinin resistance in *Plasmodium falciparum* clinical isolates. Science. 2015;347:428–31.
117. Ashley EA, Dhorda M, Fairhurst RM, Amaratunga C, Lim P, Suon S, et al. Spread of artemisinin resistance in *Plasmodium falciparum* malaria. N Engl J Med. 2014;371:411–23.
118. Conrad MD, Bigira V, Kapisi J, Muhindo M, Kamya MR, Havlir DV, et al. Polymorphisms in K13 and falcipain-2 associated with artemisinin resistance are not prevalent in *Plasmodium falciparum* isolated from Ugandan children. PLoS One. 2014;9, e105690.
119. Torrentino-Madamet M, Fall B, Benoit N, Camara C, Amalvict R, Fall M, et al. Limited polymorphisms in *k13* gene in *Plasmodium falciparum* isolates from Dakar, Senegal in 2012–2013. Malar J. 2014;13:472.
120. Ouattara A, Kone A, Adams M, Fofana B, Maiga AW, Hampton S, et al. Polymorphisms in the K13-propeller gene in artemisinin-susceptible *Plasmodium falciparum* parasites from Bougoula-Hameau and Bandiagara. Mali Am J Trop Med Hyg. 2015;92:1202–6.

121. Ménard D, Khim N, Beghain J, Adegnika AA, Shafiul-Alam M, Amodu O, et al. A worldwide map of Plasmodium falciparum K13-propeller polymorphisms. N Engl J Med. 2016;374:2453–64.
122. Miotto O, Amato R, Ashley EA, MacInnis B, Almagro-Garcia J, Amaratunga C, et al. Genetic architecture of artemisinin-resistant *Plasmodium falciparum*. Nat Genet. 2015;47:226–34.
123. Tilley L, Straimer J, Gnädig NF, Ralph SA, Fidock DA. Artemisinin action and resistance in *Plasmodium falciparum*. Trends Parasitol. 2016. doi:10.1016/j.pt.2016.05.010.
124. Mbengue A, Bhattacharjee S, Pandharkar T, Liu H, Estiu G, Stahelin RV, et al. A molecular mechanism of artemisinin resistance in *Plasmodium falciparum* malaria. Nature. 2015;520:683–7.
125. Bhattacharjee S, Stahelin RV, Speicher KD, Speicher DW, Haldar K. Endoplasmic reticulum PI(3)P lipid binding targets malaria proteins to the host cell. Cell. 2012;148:201–12.
126. Boddey JA, O'Neill MT, Lopaticki S, Carvalho TG, Hodder AN, Nebl T, et al. Export of malaria proteins requires co-translational processing of the PEXEL motif independent of phosphatidylinositol-3-phosphate binding. Nat Commun. 2016;7:10470.
127. Mok S, Ashley EA, Ferreira PE, Zhu L, Lin Z, Yeo T, et al. Population transcriptomics of human malaria parasites reveals the mechanism of artemisinin resistance. Science. 2015;347:431–5.
128. Wang J, Zhang CJ, Chia WN, Loh CC, Li Z, Lee YM, et al. Haem-activated promiscuous targeting of artemisinin in *Plasmodium falciparum*. Nat Commun. 2015;6:10111.
129. Ismail HM, Barton V, Phanchana M, Charoensutthivarakul S, Wong MH, Hemingway J, et al. Artemisinin activity-based probes identify multiple molecular targets within the asexual stage of the malaria parasites *Plasmodium falciparum* 3D7. Proc Natl Acad Sci U S A. 2016;113:2080–5.
130. Rottmann M, McNamara C, Yeung BK, Lee MC, Zou B, Russell B, et al. Spiroindolones, a potent compound class for the treatment of malaria. Science. 2010;329:1175–80.
131. Baragaña B, Hallyburton I, Lee MC, Norcross NR, Grimaldi R, Otto TD, et al. A novel multiple-stage antimalarial agent that inhibits protein synthesis. Nature. 2015;522:315–20.
132. Vincent IM, Ehmann DE, Mills SD, Perros M, Barrett MP. Untargeted metabolomics to ascertain antibiotic modes of action. Antimicrob Agents Chemother. 2016;60:2281–91.
133. Vincent IM, Creek DJ, Burgess K, Woods DJ, Burchmore RJ, Barrett MP. Untargeted metabolomics reveals a lack of synergy between nifurtimox and eflornithine against *Trypanosoma brucei*. PLoS Negl Trop Dis. 2012;6, e1618.
134. Crosnier C, Bustamante LY, Bartholdson SJ, Bei AK, Theron M, Uchikawa M, et al. Basigin is a receptor essential for erythrocyte invasion by *Plasmodium falciparum*. Nature. 2011;480:534–7.
135. Douglas AD, Williams AR, Knuepfer E, Illingworth JJ, Furze JM, Crosnier C, et al. Neutralization of *Plasmodium falciparum* merozoites by antibodies against *Pf*RH5. J Immunol. 2014;192:245–58.
136. Wright KE, Hjerrild KA, Bartlett J, Douglas AD, Jin J, Brown RE, et al. Structure of malaria invasion protein RH5 with erythrocyte basigin and blocking antibodies. Nature. 2014;515:427–30.
137. Zenonos ZA, Dummler SK, Müller-Sienerth N, Chen J, Preiser PR, Rayner JC, et al. Basigin is a druggable target for host-oriented antimalarial interventions. J Exp Med. 2015;212:1145–51.
138. Ishizuka AS, Lyke KE, DeZure A, Berry AA, Richie TL, Mendoza FH, et al. Protection against malaria at 1 year and immune correlates following PfSPZ vaccination. Nat Med. 2016;22:614–23.
139. Neafsey DE, Juraska M, Bedford T, Benkeser D, Valim C, Griggs A, et al. Genetic diversity and protective efficacy of the RTS, S/AS01 malaria vaccine. N Engl J Med. 2015;373:2025–37.
140. Wagner JC, Platt RJ, Goldfless SJ, Zhang F, Niles JC. Efficient CRISPR-Cas9-mediated genome editing in *Plasmodium falciparum*. Nat Methods. 2014;11:915–8.
141. Gantz VM, Jasinskiene N, Tatarenkova O, Fazekas A, Macias VM, Bier E, et al. Highly efficient Cas9-mediated gene drive for population modification of the malaria vector mosquito *Anopheles stephensi*. Proc Natl Acad Sci U S A. 2015;112:E6736–43.
142. Hammond A, Galizi R, Kyrou K, Simoni A, Siniscalchi C, Katsanos D, et al. A CRISPR-Cas9 gene drive system targeting female reproduction in the malaria mosquito vector *Anopheles gambiae*. Nat Biotechnol. 2016;34:78–83.
143. Callaway E. Gene-editing research in human embryos gains momentum. Nature. 2016;532:289–90.
144. Hollingdale MR, Sedegah M. Development of whole sporozoite malaria vaccines. Expert Rev Vaccines. 2016. doi:10.1080/14760584.2016.1203784.
145. Bijker EM, Borrmann S, Kappe SH, Mordmüller B, Sack BK, Khan SM. Novel approaches to whole sporozoite vaccination against malaria. Vaccine. 2015;33:7462–8.
146. Heidari R, Shaw DM, Elger BS. CRISPR and the rebirth of synthetic biology. Sci Eng Ethics. 2016. doi:10.1007/s11948-016-9768-z.
147. Pennisi E. Gene drive turns mosquitoes into malaria fighters. Science. 2015;350:1014.
148. Pearson RD, Amato R, Auburn S, Miotto O, Almagro-Garcia J, Amaratunga C, et al. Genomic analysis of local variation and recent evolution in *Plasmodium vivax*. Nat Genet. 2016. doi:10.1038/ng.3599.
149. Hupalo DN, Luo Z, Melnikov A, Sutton PL, Rogov P, Escalante A, et al. Population genomics studies identify signatures of global dispersal and drug resistance in *Plasmodium vivax*. Nat Genet. 2016. doi:10.1038/ng.3588.
150. Ghansah A, Amenga-Etego L, Amambua-Ngwa A, Andagalu B, Apinjoh T, Bouyou-Akotet M, et al. Monitoring parasite diversity for malaria elimination in sub-Saharan Africa. Science. 2014;345:1297–8.
151. Gulati S, Ekland EH, Ruggles KV, Chan RB, Jayabalasingham B, Zhou B, et al. Profiling the essential nature of lipid metabolism in asexual blood and gametocyte stages of *Plasmodium falciparum*. Cell Host Microbe. 2015;18:371–81.
152. Jones ML, Collins MO, Goulding D, Choudhary JS, Rayner JC. Analysis of protein palmitoylation reveals a pervasive role in *Plasmodium* development and pathogenesis. Cell Host Microbe. 2012;12:246–58.
153. Gomes AR, Bushell E, Schwach F, Girling G, Anar B, Quail MA, et al. A genome-scale vector resource enables high-throughput reverse genetic screening in a malaria parasite. Cell Host Microbe. 2015;17:404–13.
154. Tachibana S, Sullivan SA, Kawai S, Nakamura S, Kim HR, Goto N, et al. *Plasmodium cynomolgi* genome sequences provide insight into *Plasmodium vivax* and the monkey malaria clade. Nat Genet. 2012;44:1051–5.
155. Hall N, Karras M, Raine JD, Carlton JM, Kooij TW, Berriman M, et al. A comprehensive survey of the *Plasmodium* life cycle by genomic, transcriptomic, and proteomic analyses. Science. 2005;307:82–6.
156. Carlton JM, Angiuoli SV, Suh BB, Kooij TW, Pertea M, Silva JC, et al. Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii yoelii*. Nature. 2002;419:512–9.
157. Otto TD, Rayner JC, Böhme U, Pain A, Spottiswoode N, Sanders M, et al. Genome sequencing of chimpanzee malaria parasites reveals possible pathways of adaptation to human hosts. Nat Commun. 2014;5:4754.
158. Sundararaman SA, Plenderleith LJ, Liu W, Loy DE, Learn GH, Li Y, et al. Genomes of cryptic chimpanzee *Plasmodium* species reveal key evolutionary events leading to human malaria. Nat Commun. 2016;7:11078.
159. Philip N, Waters AP. Conditional degradation of *Plasmodium* calcineurin reveals functions in parasite colonization of both host and vector. Cell Host Microbe. 2015;18:122–31.
160. MalariaGEN Plasmodium falciparum Community Project. Genomic epidemiology of artemisinin resistant malaria. Elife. 2016;5:e08714.
161. Florens L, Washburn MP, Raine JD, Anthony RM, Grainger M, Haynes JD, et al. A proteomic view of the *Plasmodium falciparum* life cycle. Nature. 2002;419:520–6.
162. Lasonder E, Ishihama Y, Andersen JS, Vermunt AM, Pain A, Sauerwein RW, et al. Analysis of the *Plasmodium falciparum* proteome by high-accuracy mass spectrometry. Nature. 2002;419:537–42.
163. Bozdech Z, Llinás M, Pulliam BL, Wong ED, Zhu J, DeRisi JL. The transcriptome of the intraerythrocytic developmental cycle of *Plasmodium falciparum*. PLoS Biol. 2003;1, e5.
164. Doolan DL, Southwood S, Freilich DA, Sidney J, Graber NL, Shatney L, et al. Identification of *Plasmodium falciparum* antigens by antigenic analysis of genomic and proteomic data. Proc Natl Acad Sci U S A. 2003;100:9952–7.
165. Carvalho TG, Thiberge S, Sakamoto H, Menard R. Conditional mutagenesis using site-specific recombination in *Plasmodium berghei*. Proc Natl Acad Sci U S A. 2004;101:14931–6.
166. Marti M, Good RT, Rug M, Knuepfer E, Cowman AF. Targeting malaria virulence and remodeling proteins to the host erythrocyte. Science. 2004;306:1930–3.
167. Hiller NL, Bhattacharjee S, van Ooij C, Liolios K, Harrison T, Lopez-Estraño C, et al. A host-targeting signal in virulence proteins reveals a secretome in malarial infection. Science. 2004;306:1934–7.

168. LaCount DJ, Vignali M, Chettier R, Phansalkar A, Bell R, Hesselberth JR, et al. A protein interaction network of the malaria parasite *Plasmodium falciparum*. Nature. 2005;438:103–7.
169. Armstrong CM, Goldberg DE. An FKBP destabilization domain modulates protein levels in *Plasmodium falciparum*. Nat Methods. 2007;4:1007–9.
170. Shock JL, Fischer KF, DeRisi JL. Whole-genome analysis of mRNA decay in *Plasmodium falciparum* reveals a global lengthening of mRNA half-life during the intra-erythrocytic development cycle. Genome Biol. 2007;8:R134.
171. Salcedo-Amaya AM, van Driel MA, Alako BT, Trelle MB, van den Elzen AM, Cohen AM, et al. Dynamic histone H3 epigenome marking during the intraerythrocytic cycle of *Plasmodium falciparum*. Proc Natl Acad Sci U S A. 2009;106:9655–60.
172. Tewari R, Straschil U, Bateman A, Böhme U, Cherevach I, Gong P, et al. The systematic functional analysis of *Plasmodium* protein kinases identifies essential regulators of mosquito transmission. Cell Host Microbe. 2010;8:377–87.
173. van Schaijk BC, Vos MW, Janse CJ, Sauerwein RW, Khan SM. Removal of heterologous sequences from Plasmodium falciparum mutants using FLPe-recombinase. PLoS One. 2010;5, e15121.
174. O'Neill MT, Phuong T, Healer J, Richard D, Cowman AF. Gene deletion from *Plasmodium falciparum* using FLP and Cre recombinases: implications for applied site-specific recombination. Int J Parasitol. 2011;41:117–23.
175. Solyakov L, Halbert J, Alam MM, Semblat JP, Dorin-Semblat D, Reininger L, et al. Global kinomic and phospho-proteomic analyses of the human malaria parasite *Plasmodium falciparum*. Nat Commun. 2011;29:565.
176. Straimer J, Lee MCS, Lee AH, Zeitler B, Williams AE, Pearl JR, et al. Site-specific genome editing in *Plasmodium falciparum* using engineered zinc-finger nucleases. Nat Methods. 2012;9:993–8.
177. Pino P, Sebastian S, Kim EA, Bush E, Brochet M, Volkmann K, et al. A tetracycline-repressible transactivator system to study essential genes in malaria parasites. Cell Host Microbe. 2012;12:824–34.
178. Milet J, Sabbagh A, Migot-Nabias F, Luty AJ, Gaye O, Garcia A, et al. Genome-wide association study of antibody responses to *Plasmodium falciparum* candidate vaccine antigens. Genes Immun. 2016;17:110–7.
179. Otto TD, Böhme U, Jackson AP, Hunt M, Franke-Fayard B, Hoeijmakers WA, et al. A comprehensive evaluation of rodent malaria parasite genomes and gene expression. BMC Biol. 2014;12:86.
180. Pelle KG, Oh K, Buchholz K, Narasimhan V, Joice R, Milner DA, et al. Transcriptional profiling defines dynamics of parasite tissue sequestration during malaria infection. Genome Med. 2015;7:19.
181. Wei C, Xiao T, Zhang P, Wang Z, Chen X, Zhang L, et al. Deep profiling of the novel intermediate-size noncoding RNAs in intraerythrocytic *Plasmodium falciparum*. PLoS One. 2014;9, e92946.
182. Spence PJ, Jarra W, Lévy P, Reid AJ, Chappell L, Brugat T, et al. Vector transmission regulates immune control of *Plasmodium* virulence. Nature. 2013;498:228–31.
183. Ngwa CJ, Scheuermayer M, Mair GR, Kern S, Brügl T, Wirth CC, et al. Changes in the transcriptome of the malaria parasite *Plasmodium falciparum* during the initial phase of transmission from the human to the mosquito. BMC Genomics. 2013;14:256.
184. Brochet M, Collins MO, Smith TK, Thompson E, Sebastian S, Volkmann K, et al. Phosphoinositide metabolism links cGMP-dependent protein kinase G to essential Ca^{2+} signals at key decision points in the life cycle of malaria parasites. PLoS Biol. 2014;12, e1001806.
185. Alam MM, Solyakov L, Bottrill AR, Flueck C, Siddiqui FA, Singh S, et al. Phosphoproteomics reveals malaria parasite protein kinase G as a signalling hub regulating egress and invasion. Nat Commun. 2015;6:7285.
186. Tao D, Ubaida-Mohien C, Mathias DK, King JG, Pastrana-Mena R, Tripathi A, et al. Sex-partitioning of the *Plasmodium falciparum* stage V gametocyte proteome provides insight into *falciparum*-specific cell biology. Mol Cell Proteomics. 2014;13:2705–24.
187. Guttery DS, Poulin B, Ramaprasad A, Wall RJ, Ferguson DJ, Brady D, et al. Genome-wide functional analysis of *Plasmodium* protein phosphatases reveals key regulators of parasite development and differentiation. Cell Host Microbe. 2014;16:128–40.

Molecular Psychiatry (2019) 24:1576–1582
https://doi.org/10.1038/s41380-019-0445-x

PERSPECTIVE

From genetics to biology: advancing mental health research in the Genomics ERA

P. Alexander Arguello[1] · Anjené Addington[1] · Susan Borja[1] · Linda Brady[1] · Tara Dutka[1] · Miri Gitik[1] · Susan Koester[1] · Douglas Meinecke[1] · Kathleen Merikangas[1] · Francis J. McMahon[1] · David Panchision[1] · Geetha Senthil[1] · Thomas Lehner[1]

Received: 5 November 2018 / Revised: 30 April 2019 / Accepted: 3 May 2019 / Published online: 4 June 2019
© The Author(s) 2019. This article is published with open access

Abstract

The Genomics Workgroup of the National Advisory Mental Health Council (NAMHC) recently issued a set of recommendations for advancing the NIMH psychiatric genetics research program and prioritizing subsequent follow-up studies. The report emphasized the primacy of rigorous statistical support from properly designed, well-powered studies for pursuing genetic variants robustly associated with disease. Here we discuss the major points NIMH program staff consider when assessing research applications based on common and rare variants, as well as genetic syndromes, associated with psychiatric disorders. These are broad guiding principles for investigators to consider prior to submission of their applications. NIMH staff weigh these points in the context of reviewer comments, the existing literature, and current investments in related projects. Following the recommendations of the NAMHC, statistical strength and robustness of the underlying genetic discovery weighs heavily in our funding considerations as does the suitability of the proposed experimental approach. We specifically address our evaluation of applications motivated in whole, or in part, by an association between human DNA sequence variation and a disease or trait relevant to the mission of the NIMH.

Introduction

Mapping the relationship between human genetic and phenotypic variation is a challenging task for human biology. When such variation is so extreme as to disrupt normal functioning, it becomes a major challenge for medicine as well. In an effort to address this, and as part of its mission to advance mental health research, the United States' National Institute of Mental Health (NIMH) supports a broad research program in human genetics. The overarching goal is to uncover psychiatric disease mechanisms grounded in human biology.

It is clear for psychiatric disorders, and other complex disorders, that there is no simple unitary mapping between a gene and disease. For any given individual, genetic risk is a unique combination of many variants, from common to rare, that are often shared across psychiatric diagnoses. Even in the case of rare disorders and highly penetrant mutations, the cumulative genetic inheritance of the individual has a strong influence on the type and severity of the clinical presentation [1, 2]. Thus, when basic researchers ask, as they often do, "*Which gene should I study for [insert psychiatric disorder]*," the answer is not as straightforward as the question implies because it belies the underlying complexity. Indeed, the study of single genes, in isolation, may be inadequate for understanding the pathogenesis of common psychiatric disorders and should be supplanted with approaches addressing the complex interplay of genomic and other risk factors in shaping phenotypic outcome [3].

In an effort to enhance its psychiatric genetics research program and prioritize follow-up studies, the NIMH convened the Genomics Workgroup of the National Advisory

---

Supplementary information The online version of this article (https://doi.org/10.1038/s41380-019-0445-x) contains supplementary material, which is available to authorized users.

✉ P. Alexander Arguello
alexander.arguello@nih.gov

✉ Thomas Lehner
tlehner@mail.nih.gov

[1] National Institute of Mental Health, National Institutes of Health, Bethesda, MD, USA

Mental Health Council (NAMHC) comprised of geneticists and neuroscientists. Their recommendations emphasize the primacy of rigorous statistical support from properly designed, well-powered studies for pursuing genetic variants reliably associated with disease [4]. In light of these recommendations, here we provide broad guiding principles for investigators to consider when conducting studies motivated in whole, or in part, by an association between human DNA sequence variation and a psychiatric disorder or related trait. When evaluating proposed studies, we weigh these points in the context of reviewer comments, the existing literature, and current investments in related projects. Following the NAMHC report, statistical strength, and robustness of the underlying genetic discovery weighs heavily in our funding considerations as does the suitability of the proposed experimental approach. Importantly, discovery in human genetics is proceeding rapidly and genetic risk factors identified through various rigorous study designs and analytic methods across multiple cohorts and studies are more likely to stand the test of time and warrant deeper investment.

From genetic architecture

A critical lesson from the past decade of human genetics research is that regardless of whether variants are common or rare, occur in coding or noncoding regions, or alter single base pairs or large segments of the genome, strict statistical methodology is required to identify robust associations with a disease or trait. Deciding which variants to pursue for biological follow-up thus requires understanding what the initial genetic discovery does and does not tell us about the relationship to disease.

A now-common approach for genetic discovery is genome-wide association (GWA) studies. These identify statistical relationships between common single nucleotide variants across the entire genome and a phenotype of interest. Critically, they implicate regions of the genome (loci) and do not necessarily pinpoint the causal variant(s), gene(s), or mechanism(s) underlying the association. When assessing GWA findings from the literature, caution is warranted because published studies may simply list genes within risk regions, and this may give the false impression that these are the relevant disease genes. For complex traits, hundreds of regions and potentially thousands of genes may be involved that require tens to hundreds of thousands of subjects to identify [5–9]. Approaches that take into account the cumulative effects of these numerous and subtle changes in gene function may thus be more informative with regard to disease mechanisms than those focusing on individual, associated variants or genes [10].

In another approach, rare variant association studies identify genes across the genome that harbor more mutations in a group of subjects (e.g., ASD cases) compared to a well-matched group (e.g., ASD siblings) or expected by background mutation rates [11–13]. Due to limited power, sequencing studies of rare protein coding ("exome") variants often implicate genes and specific mutation categories (e.g., loss-of-function or missense mutations), but seldom individual variants. Given that each of us carry hundreds of protein altering mutations, some of which are damaging, rigorous statistical methods are required to guard against biologically interesting yet irrelevant associations [14]. Rare variant discovery methods differ in their assumptions, power, error rates, and how they leverage genetic and nongenetic information to identify genome-wide statistical associations [15–18]. With any approach, a highly rigorous statistical framework is essential [19]. In some cases, more liberal statistical thresholds may be appropriate to identify larger sets of genes to be interrogated collectively, either bioinformatically or experimentally, where the biological signal prevails over the noise. With limited resources, however, a conservative approach relies on the strictest statistical criteria to select genes of high confidence for deep, targeted experimental follow-up (see Supplemental information).

In some cases, symptoms of rare diseases and syndromes may overlap with those of more common and etiologically complex disorders. For example, individuals with Fragile X or velocardiofacial syndromes often have features of autism or schizophrenia, respectively. Although these syndromes have well-defined genetic causes, the manifestation of different symptoms is likely influenced by other genetic factors. Thus, mutations in *FRM1* and copy number variants (CNVs) of 22q11 that cause these syndromes should not be considered synonymous with autism or schizophrenia. In fact, the more common neurological feature for both these syndromes is intellectual disability [20, 21]. Even in the case of fairly penetrant variants such as CNVs, unbiased genetic associations are required to establish a definitive link between these variants and a psychiatric diagnosis [22]. This holds true for individual genes within large, multigenic CNVs. Currently, there is no robust statistical genetic evidence that these CNVs can be resolved to individual risk genes [17] nor are there consensus functional experiments for determining which gene or genes within or near an associated CNV is driving disease risk.

Overall, there are many factors to consider when assessing the strength of evidence for a genetic association and it may be difficult without familiarity of statistical or quantitative genetics (see Supplemental information for an FAQ). We recognize that there are still only a handful of causal variants and genes with a high level of statistical confidence across psychiatric disorders, but we expect this number to grow with increases in sample size, better functional annotation and improved methods for causal inference [23].

Box 1. Points to consider when experimentally following up human genetic findings

A. Assessing the strength of the evidence for genetic association with a disease or trait

- Is the genetic association statistically robust using an unbiased, genome-wide approach (e.g., genome-wide significant)? Has it been replicated? What is the effect size?
- For large, multigenic CNVs, have any individual gene(s) been statistically associated genome wide?
- Has the genome-wide significant signal been statistically narrowed down (i.e., fine mapped) to a causal variant, gene, and/or isoform(s)?
- What other diseases or traits has the variant or gene been associated with? How does that inform the design and interpretation of the proposed experimental study?

B. Experimental study design based on the variant(s) or gene(s) of interest

- Has the functional variant(s) been identified? Does it cause a gain or loss of function? Does it increase or decrease gene expression? Splicing? Transcription factor binding?
- In what human tissues and cell types is that variant active? During which developmental period(s)?
- How conserved is the implicated genomic element or corresponding gene(s) across species? How about the regulatory regions? Or the gene's coexpression and/or protein interaction network across cells, tissues, and developmental periods?
- Do genome-wide enrichment analyses for the trait or disease identify those same cells, tissues, and developmental periods?
- Is the function of the gene, isoform, or effect of the variant robust across multiple cell lines and animal strains?
- Is the proposed experimental system appropriate given the conservation of the genomic element, gene, or isoform and its expression across cell, tissues, developmental periods, and/or species?
- If studying risk variants or genes individually or collectively, is there an appropriate set of well-matched control variants or genes?

C. Maximizing utility and impact of the approach

- Is the proposed experimental approach best suited to address the mechanism of disease risk conferred by a particular genetic variant or to investigate the basic biology and function of the variant and corresponding gene/isoform(s)?
- Given the question of interest, what is the trade-off between a deep dive into the biology and function of a particular gene or variant versus a comprehensive assessment of all genome-wide significant, disease or trait-associated genes or variants?
- Have the experimental effects of genes or variants been considered in the context of all clinical and nonclinical traits associated with those same variants as well as whether nonassociated genes or variants have similar effects?

We provide some suggestions (Box 1a) for prioritizing where to focus experimental efforts to gain relevant biological traction.

To biology

Once a reliable association is identified, the crucial and more difficult step is transforming that genetic discovery into biological insight. Although true genetic associations must, by necessity, signal a causal effect, that effect may be indirect and far removed mechanistically from the trait of interest [24]. This has important implications for the selection, design, and interpretation of experimental paradigms. Psychiatric disorders lack specific markers and genetic variants often influence many traits (i.e., pleiotropy) [25]. This reality limits the ability to draw causal inferences about disease mechanisms. It is thus critical to clearly define the research question and carefully select the most appropriate experimental system to address that question (Box 1b). To illustrate these points, we provide and evaluate five examples of hypothetical studies related to mental health aiming to biologically follow-up genetic findings.

(1) Embracing pleiotropy and polygenicity

> *Analyses of well-powered GWA and exome studies show that implicated genes as a whole are preferentially expressed in region X of the human brain relative to other regions. Region X consists of a mixed population of cells that project to different downstream brain regions. Our preliminary data indicate that well-powered GWA and sequencing studies from related diseases also show an enrichment, albeit weaker one, in region X. Our data further indicate an enrichment of orthologous genes in brain region X of species Y. Here we use our method of isolating cells*

> *based on their projection targets in species Y to better resolve the genetic enrichment observed across these diseases into projection-specific cell types. We will further develop transcriptional signatures of these cells and validate our findings in human brain. This study will increase our understanding of how genetic risk across related diseases map onto specific cells and circuits in the brain.*

This proposed study does not focus on any one gene or gene variant, but rather follows up on well-powered analyses that take into account many potential genes associated with disease. Furthermore, given the known overlap in the causes and symptoms of psychiatric disorders, it does not focus on any one disease. It takes a finding based on human genetics and human tissue, identifies a similar biological signal in another species, then leverages that experimental system to further refine that signal and validate it back in humans. Using genetic information in this manner may reveal common biological mechanisms that would otherwise be missed by single gene and single disease studies.

(2) Getting back to basics

> *Gene X has been fine-mapped as a likely causal gene within a genome-wide significant disease risk locus and replicated across multiple cohorts. (Alternatively, gene X has a genome-wide significant increased* mutation *burden in cases versus controls across multiple cohorts). Here we propose to further understand the function of gene X in the brain. Our preliminary data show that in both an experimental organism and humans gene X is most highly expressed in a specific cell type known to be important for a particular process. We will transcriptionally profile this cell type across development in the organism to identify the time course of gene X expression and determine if it coincides with the maturation of the relevant cellular process. We will also transcriptionally profile wildtype and conditional, cell-type selective gene X knockout cells, assess the impact on the cellular process and identify potential molecular mechanisms via rescue of differentially expressed targets. This study will advance our understanding of gene X biology.*

Although this proposed study is based on a statistically robust genetic finding, it is intended as a basic science study into the biology of gene X. The strength of the study depends on how much and how well it will advance our understanding of gene X and/or fundamental principles of biology. These findings may or may not be relevant for understanding how human genetic variation in gene X contributes to variation in disease risk, but it will contribute basic knowledge whose ultimate impact will be determined in the future.

(3) Updating poor priors

> *A recent genome-wide association study identified 12p34 as a genome-wide significant disease risk locus. Interestingly, one of the three genes within the risk locus is gene X that is part of a pathway long hypothesized to be involved in disease pathogenesis. How gene X is involved in this pathway is still not fully understood. Our preliminary data show that manipulating expression of gene X produces changes in the pathway similar to those observed in patients. Here we will use a series of loss- and gain-of-function experiments to characterize how the function of gene X influences the function of the pathway. This study will advance our understanding of gene X and how it contributes to disease.*

This proposed study is based on the location of gene X within a genome-wide significant risk region. There is, however, no evidence that the association with disease is operating via that gene as opposed to the other two genes within the region or possibly other more distant genes. Such evidence would require a statistical and/or functional fine-mapping procedure [26, 27]. The information that gene X is part of a disease-relevant pathway may appear to provide a strong prior for focusing on that particular gene. Yet, in the absence of a formal unbiased assessment of how many pathways have been associated with disease, how many genes are within each of those pathways, how these genes are distributed across the genome, or how many risk regions and genes overlap with those regions, it is not possible to determine whether this is a chance occurrence or not. There may be other lines of evidence that indicate gene X and its pathway are important to study in terms of fundamental biology, but in this example the evidence of disease relevance is circumstantial and may in fact detract from the potential basic science value of the study.

(4) Studying rare variants

> *Gene X has a genome-wide significant increased burden of heterozygous loss-of-function mutations in cases versus controls across multiple cohorts using various analytic methods. In order to understand disease pathogenesis and pathophysiology, here we propose to characterize the effects of gene X knockout at the molecular, cellular, and behavioral levels in an experimentally tractable organism. Our preliminary data show disease-like deficits in behavior that are*

rescued by expressing gene X in adult animals. This study will identify the molecular and cellular mechanisms by which gene X causes disease and identify potential therapeutic targets.

Although this proposed study is based on a statistically robust genetic finding, there are limitations to the experimental approach. The application proposes characterization of an animal knockout (null allele) that may or may not mimic heterozygous loss-of-function mutations observed in patients. The goal is to explore disease pathogenesis and pathophysiology and not the basic biology of the gene, but the proposal assumes that because mutants exhibit "disease-like" behavioral deficits, changes at the molecular, or cellular level will be de facto relevant for understanding human disease mechanisms. Such changes must be interpreted in a wider biological and clinical context. For example, how many null mutations in other genes not associated with the disease cause similar changes in the organism? This problem may be mitigated by collectively studying multiple disease-associated genes to look for common patterns of effects. Yet, that design would require an appropriate set of well-matched control genes assessed for similar changes. Critically, "disease-like" behavior in an animal does not establish a causal link among a gene, molecular/cellular effect, and disease nor does the absence of such behavior refute it.

It is important to keep in mind that loss-of-function mutations in particular genes are often associated with multiple neurological and psychiatric phenotypes in people, are not fully penetrant and are modified by polygenic effects [2, 28]. This lack of specificity at both the genetic and phenotypic levels should guide and constrain the interpretation of phenotypes observed in mutant organisms. In general, broad characterization of mutant animals are less informative than those that first take into account wider biological and clinical factors to formulate well-defined research questions (Box 1).

(5) Studying common variants

Gene X has been fine-mapped as a likely causal gene within a genome-wide significant disease risk locus and replicated across multiple cohorts. In order to understand disease pathogenesis and pathophysiology, here we propose to characterize the effects of gene X knockout at the molecular, cellular, and behavioral levels in an experimentally tractable organism. Our preliminary data show disease-like deficits in behavior that are rescued by expressing gene X in adult animals. This study will identify the molecular and cellular mechanisms by which the gene causes disease and identify potential therapeutic targets.

This proposed study is similar to the previous example except that it is based on a common variant association mapped to a gene. In addition to the limitations noted above, the approach has the added issue of attempting to recapitulate, with a null allele, what are often subtle, cell-type specific regulatory effects of common noncoding variants. These risk variants in humans are typically of low effect and operate in the context of hundreds of other such variants. The biological effect of a null allele in an experimental system is thus not necessarily related to a single risk variant in the context of this polygenic effect in humans. It is for this reason and those mentioned above that studies attempting to create a genetic 'model of disease' are potentially more problematic than those focused on either the basic biology of likely causal variants, or their associated genes, within a risk locus or a more integrative approach that accounts for a multitude of risk variants [29].

Summary

The recent and rapid progress in psychiatric genetics has simultaneously created a set of opportunities and challenges (Box 1c). Previously, when genetic clues were scarce, investments in experimental studies tolerated potential false alarms to avoid possible misses for much needed hits. Now, however, with a steady stream of reliable genetic findings, false positives have the danger of diverting precious resources. The availability of large-scale, multidimensional genetic, and genomic datasets combined with incentivized motivated reasoning, confirmation bias, and apophenia [30] poses the danger of generating appealing but unsupported biological narratives. An emphasis on the strength of statistical support for genetic discoveries serves as an important check and a guide for resource investment. This does not imply that genetics are not informed by prior knowledge and biological data, but such information needs to be formally and rigorously incorporated into the whole body of evidence [14, 31, 32], not applied in an ad hoc manner—it is important to systematically integrate statistical and biological evidence to reduce the risk of introducing bias. We hope that such an approach, coupled with appropriate experimental systems, will at most reveal disease-relevant mechanisms and at least advance our understanding of fundamental biology.

Acknowledgements We wish to express our gratitude to the members of the NIMH Genomics Team and Drs Josh Gordon and Naomi Wray for their helpful comments and suggestions. We thank Lora Bingaman for her tireless support of the NIMH Genomics Team and its deliberations and for her editorial assistance with this manuscript.

Compliance with ethical standards

Conflict of interest The authors report no biomedical financial interests or potential conflicts of interest. The views and opinions expressed in this article are those of the authors and should not be construed to represent the views of any of the sponsoring organization, agencies, or the US government.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Stefansson H, Meyer-Lindenberg A, Steinberg S, Magnusdottir B, Morgen K, Arnarsdottir S, et al. CNVs conferring risk of autism or schizophrenia affect cognition in controls. Nature. 2014;505:361–366.
2. Niemi MEK, Martin HC, Rice DL, Gallone G, Gordon S, Kelemen M, et al. Common genetic variants contribute to risk of rare severe neurodevelopmental disorders. Nature. 2018;268–271.
3. Hyman SE. The daunting polygenicity of mental illness: making a new map. Philos Trans R Soc Lond B Biol Sci. 2018;373:1–8.
4. Report of the National Advisory Mental Health Council Workgroup on Genomics: opportunities and challenges of psychiatric genetics 2017. https://www.nimh.nih.gov/about/advisory-boards-andgroups/namhc/reports/report-of-the-national-advisory-mental-health-council-workgroup-ongenomics.shtml.
5. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. 10 years of GWAS discovery: biology, function, and translation. Am J Hum Genet. 2017;101:5–22.
6. Grove J, Ripke S, Als TD, Mattheisen M, Walters RK, Won H, et al. Identification of common genetic risk variants for autism spectrum disorder. Nat Genet. 2019;51:431–444.
7. Pardinas AF, Holmans P, Pocklington AJ, Escott-Price V, Ripke S, Carrera N, et al. Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. Nat Genet. 2018;50:381–389.
8. Wray NR, Ripke S, Mattheisen M, Trzaskowski M, Byrne EM, Abdellaoui A, et al. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. Nat Genet. 2018;50:668–681.
9. Demontis D, Walters RK, Martin J, Mattheisen M, Als TD, Agerbo E, et al. Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. Nat Genet. 2019;51:63–75.
10. Skene NG, Bryois J, Bakken TE, Breen G, Crowley JJ, Gaspar HA, et al. Genetic identification of brain cell types underlying schizophrenia. Nat Genet. 2018;50:825–833.
11. Kosmicki JA, Samocha KE, Howrigan DP, Sanders SJ, Slowikowski K, Lek M, et al. Refining the role of de novo protein-truncating variants in neurodevelopmental disorders by using population reference samples. Nat Genet. 2017;49:504–510.
12. Singh T, Kurki MI, Curtis D, Purcell SM, Crooks L, McRae J, et al. Rare loss-of-function variants in SETD1A are associated with schizophrenia and developmental disorders. Nat Neurosci. 2016;19:571–577.
13. Genovese G, Fromer M, Stahl EA, Ruderfer DM, Chambert K, Landen M, et al. Increased burden of ultra-rare protein-altering variants among 4,877 individuals with schizophrenia. Nat Neurosci. 2016;19:1433–1441.
14. MacArthur DG, Manolio TA, Dimmock DP, Rehm HL, Shendure J, Abecasis GR, et al. Guidelines for investigating causality of sequence variants in human disease. Nature. 2014;508:469–476.
15. Deciphering Developmental Disorders S. Prevalence and architecture of de novo mutations in developmental disorders. Nature. 2017;542:433–438.
16. Krishnan A, Zhang R, Yao V, Theesfeld CL, Wong AK, Tadych A, et al. Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. Nat Neurosci. 2016;19:1454–1462.
17. Sanders SJ, He X, Willsey AJ, Ercan-Sencicek AG, Samocha KE, Cicek AE, et al. Insights into autism spectrum disorder genomic architecture and biology from 71 risk loci. Neuron. 2015;87:1215–1233.
18. Werling DM, Brand H, An JY, Stone MR, Zhu L, Glessner JT. et al. An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder. Nat Genet2018;50():727–736.
19. Satterstrom FK, Kosmicki JA, Wang J, Breen MS, De Rubeis S, An J-Y, et al. Novel genes for autism implicate both excitatory and inhibitory cell lineages in risk. 2018. bioRxiv 484113.
20. Hoeffding LK, Trabjerg BB, Olsen L, Mazin W, Sparso T, Vangkilde A, et al. Risk of psychiatric disorders among individuals with the 22q11.2 deletion or duplication: a Danish nationwide, register-based study. JAMA Psychiatry. 2017;74:282–290.
21. Sherman SL, Kidd SA, Riley C, Berry-Kravis E, Andrews HF, Miller RM, et al. FORWARD: a registry and longitudinal clinical database to study fragile X syndrome. Pediatrics. 2017;139(Suppl 3):S183–S193.
22. Marshall CR, Howrigan DP, Merico D, Thiruvahindrapuram B, Wu W, Greer DS, et al. Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. Nat Genet. 2017;49:27–35.
23. Gandal MJ, Zhang P, Hadjimichael E, Walker RL, Chen C, Liu S, et al. Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. Science. 2018;362:6420.
24. Thorgeirsson TE, Geller F, Sulem P, Rafnar T, Wiste A, Magnusson KP, et al. A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. Nature. 2008;452:638–642.
25. Brainstorm C, Anttila V, Bulik-Sullivan B, Finucane HK, Walters RK, Bras J, et al. Analysis of shared heritability in common disorders of the brain. Science. 2018;360:6395.
26. Schaid DJ, Chen W, Larson NB. From genome-wide associations to candidate causal variants by statistical fine-mapping. Nat Rev Genet. 2018;19:491–504.

27. Pasaniuc B, Price AL. Dissecting the genetics of complex traits using summary association statistics. Nat Rev Genet. 2017;18: 117–127.
28. Weiner DJ, Wigdor EM, Ripke S, Walters RK, Kosmicki JA, Grove J, et al. Polygenic transmission disequilibrium confirms that common and rare variation act additively to create risk for autism spectrum disorders. Nat Genet. 2017;49:978–985.
29. Wray NR, Wijmenga C, Sullivan PF, Yang J, Visscher PM. Common disease is more complex than implied by the core gene omnigenic model. Cell. 2018;173(7):1573–1580.
30. Munafò MR, Nosek BA, Bishop DVM, Button KS, Chambers CD, Percie du Sert N, et al. A manifesto for reproducible science. Nat Hum Behav. 2017;1:1–9.
31. Samocha KE, Robinson EB, Sanders SJ, Stevens C, Sabo A, McGrath LM, et al. A framework for the interpretation of de novo mutation in human disease. Nat Genet. 2014;46:944–950.
32. Zhou J, Theesfeld CL, Yao K, Chen KM, Wong AK, Troyanskaya OG. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. Nat Genet. 2018;50:1171–1179.